

SCHAUM'S
ouTlines

全美经典 学习指导系列

统计学

(第三版)

[美] M. R. 斯皮格尔 L. J. 斯蒂芬斯 著

杨纪龙 杜秀丽 姚奕 赵媛媛 梁志彬 译

508道详细解答的习题

694道附加习题

包括几百道习题的计算机解

优秀教科书的有益补充

自学的良师益友



科学出版社

麦格劳-希尔教育出版集团

(O-1515.0101)

责任编辑: 毕 颖

全球销量
超越 的

SCHAUM'S
ouTlines

“全美经典学习指导系列” 是您的最佳 学习伴侣!

40年来最畅销的教辅系列

全美著名高校资深教授倾力之作

国内重点高校任课教师全力推荐并担当翻译

省时高效的学习辅导, 全面详细的习题解答

迄今为止国内最全面的教辅系列

覆盖大学理工科专业

全美经典学习指导系列

概率和统计

统计学

离散数学

Mathematica使用指南

数理金融引论

机械振动

微分方程

统计学原理(上)

统计学原理(下)

微积分

静力学与材料力学

有限元分析

传热学

近代物理学

2000工程力学习题精解

工程力学

3000物理习题精解

流体动力学

物理学基础

材料力学

2000离散数学习题精解

工程热力学

数值分析

量子力学

有机化学习题精解

3000化学习题精解

大学化学习题精解

电路

电气工程基础

工程电磁场基础

数字信号处理

数字系统导论

数字原理

电机与机电学

基本电路分析

信号与系统

微生物学

生物化学

生物学

分子和细胞生物学

人体解剖与生理学

<http://www.sciencep.com>

<http://www.mheducation.com>

ISBN 7-03-009620-7



9 787030 096203 >

Mc
Graw
Hill

ISBN 7-03-009620-7/O · 1515

定价: 35.00 元

全美经典学习指导系列

统 计 学

(第 三 版)

[美] M. R. 斯皮格尔 L. J. 斯蒂芬斯 著

杨纪龙 杜秀丽 姚奕 赵媛媛 梁志彬 译

科 学 出 版 社

麦格劳-希尔教育出版集团

2 0 0 2

内 容 简 介

本书在对主要的数学概念进行了回顾后,清晰地讲述了数学和其他科学领域都会需要的统计学的基础知识,从变量和图表到标准分布再到基本的概率和样本理论.介绍了用最流行的统计软件包来解决问题.书中众多的例题及详解可进一步强化对理论和方法的理解.

本书可供大学数学、统计学等专业的教师、学生以及各类专业技术人员参考.

Murray R. Spiegel, Larry J. Stephens, Schaum's Outline of Theory and Problems of Statistics, Third Edition.

ISBN:0-07-060281-6

Copyright © 1999 by the McGraw-Hill Companies, Inc.

Authorized translation from the English language edition published by McGraw-Hill Companies, Inc.

All rights reserved.

本书中文简体字版由科学出版社和美国麦格劳-希尔教育出版集团合作出版.未经出版者书面许可,不得以任何方式复制或抄袭本书的任何部分.

版权所有,翻印必究.

本书封面贴有 McGraw-Hill 公司防伪标签,无标签者不得销售.

图字:01-2001-1767 号

图书在版编目(CIP)数据

统计学(第三版)/[美]斯皮格尔(Spiegel, M. R.)等著,杨纪龙等译. - 北京:科学出版社,2002

(全美经典学习指导系列)

ISBN 7-03-009620-7

I. 统… II. ①斯…②杨… III. 统计学 IV. C8

中国版本图书馆 CIP 数据核字(2001)第 047881 号

科学出版社 出版

北京东黄城根北街16号

邮政编码:100717

<http://www.sciencep.com>

双青印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2002 年 1 月第 一 版 开本:A4(890×1240)

2002 年 1 月第一次印刷 印张:25 3/4

印数:1—5 000 字数:732 000

定价:35.00 元

(如有印装质量问题,我社负责调换〈环伟〉)

译者的话

本书是美国著名统计学家 M.R. 斯皮格尔和 L.J. 斯蒂芬斯所著的一本统计学教材,系《全
美经典学习指导系列》丛书之一.主要内容有频数分布,抽样理论,估计理论,假设检验,方差分
析,回归和相关分析,时间序列分析及过程统计控制等.

本书有如下特色:1. 阅读本书所需的数学基础仅是算术和初等代数学的知识,但本书所
收内容在深度和广度上都超过一般的统计学初级教程.2. 书中附有大量的已给出详细解答的
各类实际应用问题,它们占有全书的大部分篇幅.3. 对各种统计方法都介绍了统计软件包的
使用.这些特色在国内的统计学教材中尚不多见.本书是一本优秀的统计学教材,可供大学数
学、统计学等专业师生及各类专业技术人员参考.

参加本书翻译的有南京师范大学数学与计算机科学学院的杨纪龙、杜秀丽、姚奕、赵媛媛
和梁志彬.其中姚奕译第一至第五章,赵媛媛译第六至第十章,梁志彬译第十一至十五章,杜
秀丽译第十六至第十九章,杨纪龙修改整理全部译稿.在翻译中我们对原著中的若干错误作了
改正,不在书中一一指出.由于时间仓促,翻译中可能出现错误,欢迎广大读者给予指正.

在翻译本书过程中,科学出版社的毕颖编辑提出了许多宝贵的建议,我们深表感谢.

译者

2001 年 10 月

第三版序言

自从 1961 年第一版出版后,工业技术和社会科学都发生了巨大的变化,在第三版中我们修改了过时的内容.例如,在第二版中的无线电真空管问题,由于 30 岁以下的人大多不知无线电真空管为何物,这个问题以及许多其他问题已经换为诸如健康、艾滋病、英特网技术、电话等问题.数学和统计学的面貌没有变,而应用的领域以及统计的计算方面改变了.

另一个重要的改进是在文中对统计软件进行了介绍.SAS, SPSS 和 MINITAB 等统计软件的发展极大地推动了统计软件在现实生活问题中的应用. MINITAB 软件包是应用最为广泛的统计软件包之一.我要感谢 Minitab Inc. 允许我在本书中使用 MINITAB 的产品.如今许多现代统计学教科书都把计算机软件作为课本内容的一部分,我之所以选择 MINITAB,是因为它应用广泛且使用方便.学生学会了使用 MINITAB 的各种数据文件结构和命令及子命令的结构,也就能使用别的统计软件了.有了下拉式菜单和对话方式,这一软件使用就变得更加方便.我们在书中涉及 MINITAB 的部分讨论了该软件的各种命令和下拉菜单的使用.

本书许多新问题的讨论都涉及到了重要的统计学概念—— p -值检验.当 1961 年第一版出版时,由于没有计算机软件的辅助,它的应用远不如今天.如今 P -值的计算在计算机软件中只是平常的事,因此它已经由统计软件包相应提供.

第一版中的第十九章“指数”已由新的一章“过程统计控制和过程性能”所代替.这一章中讨论的许多问题都在工业上有应用,因此我觉得有必要将它们引入本书.统计软件包中介绍的过程统计控制和过程性能的技术为把这些技术应用到许多工业背景中去提供了方便.软件完成了所有的计算工作,这些计算相当繁琐.我选择使用 MINITAB 是因为我觉得它在 SPC 应用中是最好的.

我要感谢我的妻子 Lana 对我工作的理解,感谢我的朋友 Stanley Wileman 在计算机方面给我提供的帮助,还有 Alan Hunt 和英国伦敦 Keyword 出版有限公司的所有工作人员,感谢他们的杰出工作.最后,我要感谢 McGraw-Hill 工作人员的合作和帮助.

L. J. 斯蒂芬斯

第二版序言

统计或者说统计方法,在人类几乎所有的活动领域中扮演着越来越重要的角色.原来因处理国家事务而得名的统计,现在已影响到农业、生物、商业、化工、通信、经济、教育、电子、医药、物理、政治学、心理学、社会学以及许多其他科学和工程领域.

本书旨在介绍一般的统计学原理,这些原理可能对所有人都会有所帮助,而不管他们专门研究哪个方面.本书可作为所有现行标准教材的补充,还可作为统计学的正规教材.对于那些正在致力于把统计应用到他们所研究的领域中的人,本书也应有相当大的帮助.

本书每章开始都用图例和其他描述性材料清楚地阐述了有关的定义、定理和法则,然后分类给出已给解答的和补充的问题,这些问题大多来自实际应用.其中已经给出解答的问题可以阐明原理,增强读者对本书的理解.这些问题突出了一些重点原理,使得学生不会在学习过程中感到没有依据.这些问题还不断重复了对教学效率至关重要的基本原则,包括许多公式的衍生式.大量的附答案的补充问题,可以使学习者对每章所学的内容进行全面的复习.

理解本书仅需的数学背景是算术和代数学的知识.本书第一章复习了重要的数学概念,这一章既可以在课程开始的时候学习,也可以在今后需要时学习和参考.

本书的前几章涉及的问题是频数分布和相关的中心趋势、离差、偏度和峰度.然后导入对基本概率原理和应用的讨论,为学习抽样理论做好铺垫.抽样理论首先涉及与正态分布相关的大样本理论.统计估计和假设检验及显著性问题.其后一章讨论与小样本理论相关的学生氏 t 分布, χ^2 分布, F 分布及应用.曲线拟合和最小二乘法这一章介绍了与两个变量有关的相关性和回归问题.其后的章节涉及到方差分析和非参数方法,这两个问题都是第一次在第二版中谈到.最后两章分别讨论了时间序列分析和指数问题.

本书所收内容比大多数初级课程涉及的内容多得多,这样使用起来也就更灵活,既可以是一本很有用的参考书,也可以激发读者对本书讨论的问题产生更大的兴趣.使用本书时,可以不必按其编排的顺序学习,甚至可以略过某些章节而不会给读者带来任何困难.例如,如果希望在学习抽样理论前,先学习相关分析,回归分析,时间序列和指数问题,可以在第五章之后直接学习第十三到第十五章和第十八到第十九章.类似地,如果不想花太多的时间学习概率,可以省略第六章的大部分内容作为基础课程,第十五章的所有内容可以不学.我们按照现在的顺序编排本书,是因为在现代课程中,越来越趋向于尽早介绍抽样理论和统计影响的问题.

我要感谢很多政府和私人的机构为本书的表格提供数据,为此,全书表格中的数据都注明了其来源.在此,我要特别感谢英国皇家学会会员剑桥大学教授 Ronald A. Fisher 爵士、英国皇家学会会员洛桑(Rothamsted)的 Frank Yates 博士和爱丁堡的 Messrs. Oliver and Boyd Ltd 公司,感谢他们允许本书使用《生物、农业和医药研究统计用表》一书中表格 3 的数据.我还要感谢 Esther 和 Meyer Scher 的鼓励以及 McGraw-Hill 公司工作人员的合作.

M. R. 斯皮格尔

目 录

第一章 变量和图形	(1)
统计学	(1)
总体和样本, 归纳统计学和描述性统计学	(1)
离散变量和连续变量	(1)
数据舍入	(1)
科学记数	(2)
有效数字	(2)
数值计算	(3)
函数	(3)
直角坐标	(3)
图形	(4)
方程	(4)
不等式	(4)
对数	(5)
反对数	(5)
对数计算	(6)
第二章 频数分布	(28)
原始数据	(28)
数组阵列	(28)
频数分布	(28)
组距和组限	(28)
组界	(28)
组距的大小或宽度	(29)
组中值	(29)
建立频数分布的一般法则	(29)
直方图和频数多边形	(29)
频率分布	(29)
累积频数分布和卵形线	(30)
累积频率分布和百分率卵形线	(30)
频数曲线和光滑卵形线	(31)
频数曲线的种类	(31)
第三章 均值, 中位数, 众数以及其他表示集中趋势的度量	(45)
下标, 记法	(45)
求和符号	(45)
平均值或集中趋势的度量	(45)
算术平均	(45)
加权算术平均	(46)
算术平均的性质	(46)
从分类资料中计算算术平均值	(46)
中位数	(47)

众数	(47)
均值, 中位数和众数间的经验关系	(48)
几何平均 G	(48)
调和平均 H	(48)
算术平均, 几何平均和调和平均间的关系	(48)
均方根(RMS)	(49)
四分位数, 十分位数和百分位数	(49)
第四章 标准差和其他表示离差的度量	(67)
离差或变差	(67)
全距	(67)
平均偏差	(67)
半内四分位数间距	(67)
10~90 百分位数间距	(68)
标准差	(68)
方差	(68)
计算标准差的快捷方法	(68)
标准差的性质	(69)
Charlier 检验	(69)
Sheppard 方差修正	(70)
离差度量间的经验关系	(70)
绝对和相对离差, 变异系数	(70)
标准化变量, 标准分数	(70)
第五章 矩, 偏度和峰度	(86)
矩	(86)
分类资料的矩	(86)
矩间关系	(86)
分类资料矩的计算	(86)
Charlier 检验和 Sheppard 修正	(87)
无量纲形式的矩	(87)
偏度	(87)
峰度	(88)
总体矩, 偏度和峰度	(88)
第六章 初等概率论	(96)
概率的定义	(96)
条件概率, 独立和不独立事件	(96)
互不相容事件	(97)
概率分布	(98)
数学期望	(99)
总体均值和方差与样本均值和方差的关系	(99)
组合分析	(99)
组合	(100)
$n!$ 的 Stirling 逼近	(100)
概率和集合论的关系	(100)
第七章 二项分布, 正态分布和泊松分布	(117)
二项分布	(117)

正态分布	(118)
二项分布和正态分布的关系	(119)
泊松分布	(119)
二项分布和泊松分布的关系	(119)
多项分布	(119)
用样本的频率分布拟合理论分布	(120)
第八章 初等抽样理论	(137)
抽样理论	(137)
随机样本和随机数	(137)
有放回和无放回抽样	(137)
抽样分布	(137)
均值的抽样分布	(138)
比例的抽样分布	(138)
差与和的抽样分布	(138)
标准误差	(139)
第九章 统计估计理论	(152)
参数的估计	(152)
无偏估计	(152)
有效估计	(152)
点估计和区间估计	(152)
总体参数的置信区间估计	(153)
可能误差	(154)
第十章 统计决策理论	(162)
统计决策	(162)
统计假设	(162)
假设检验, 显著性检验或决策法则	(162)
第一类和第二类错误	(162)
显著性水平	(162)
关于正态分布的检验	(163)
双边检验和单边检验	(163)
特殊检验	(164)
OC 曲线, 检验的功效	(164)
控制图	(164)
有关样本差的检验	(164)
关于二项分布的检验	(165)
第十一章 小样本理论	(180)
小样本	(180)
t 分布	(180)
置信区间	(181)
假设检验和显著性检验	(181)
χ^2 分布	(182)
χ^2 的置信区间	(182)
自由度	(183)
F 分布	(183)

第十二章 χ^2 检验	(194)
观察频数和理论频数	(194)
χ^2 的定义	(194)
显著性检验	(194)
拟合优度的 χ^2 检验	(195)
列联表	(195)
关于连续性的 Yates 修正	(195)
计算 χ^2 的简单公式	(195)
列联系数	(196)
属性相关	(196)
χ^2 的可加性	(196)
第十三章 曲线拟合和最小二乘法	(209)
变量间的相互关系	(209)
曲线拟合	(209)
近似曲线的方程	(209)
曲线拟合的徒手法	(210)
直线	(210)
最小二乘法	(210)
最小二乘直线	(211)
非线性关系	(212)
最小二乘抛物线	(212)
回归	(212)
时间序列的应用	(212)
两个以上变量的问题	(212)
第十四章 相关理论	(232)
相关与回归	(232)
线性相关	(232)
相关性度量	(232)
最小二乘回归直线	(233)
估计的标准误差	(233)
回归平方和与残差平方和	(234)
相关系数	(234)
关于相关系数的附注	(234)
线性相关系数的积-矩公式	(235)
快捷计算公式	(235)
回归直线和线性相关系数	(236)
时间序列相关	(236)
属性相关	(236)
相关的抽样理论	(236)
回归的抽样理论	(237)
第十五章 多重相关与偏相关	(258)
多重相关	(258)
下标记号	(258)
回归方程和回归平面	(258)
最小二乘回归平面的正规方程	(258)

回归平面和相关系数	(259)
估计的标准误差	(259)
多重相关系数	(259)
因变量的转换	(260)
多于三个变量的推广	(260)
偏相关	(260)
多重相关系数与偏相关系数之间的关系	(260)
非线性多重回归	(261)
第十六章 方差分析	(271)
方差分析的目的	(271)
单向分类或单因素试验	(271)
总变差, 组内变差和组间变差	(271)
计算变差的快捷方法	(272)
方差分析的数学模型	(272)
变差的数学期望	(272)
变差的分布	(273)
等均值零假设的 F 检验	(273)
方差分析表	(273)
观测值数目不等时所做的修正	(274)
双向分类或双因素试验	(274)
双因素试验的记号表示	(274)
双因素试验的变差	(275)
双因素方差分析	(275)
有重复的双因素试验	(277)
实验设计	(278)
第十七章 非参数检验	(302)
引言	
符号检验	(302)
Mann-Whitney U 检验	(302)
Kruskal-Wallis H 检验	(303)
有结点时 H 检验的修正	(304)
随机性的游程检验	(304)
游程检验的进一步应用	(304)
Spearman 秩相关	(305)
第十八章 时间序列分析	(326)
时间序列	(326)
时间序列图	(326)
时间序列的特征运动	(326)
时间序列运动分类	(326)
时间序列分析	(327)
移动平均, 时间序列的平稳化	(327)
趋势的估计	(328)
季节变差的估计, 季节指数	(328)
数据的消季节化	(329)
循环变差的估计	(329)

不规则变差的估计	(329)
数据的可比性	(329)
预测	(329)
时间序列分析的基本步骤小结	(329)
第十九章 过程统计控制和过程性能	(353)
对控制图的一般讨论	(353)
变量和属性控制图	(353)
$\bar{X} - R$ 图	(354)
指定原因的检验	(355)
过程性能	(356)
P - 图和 NP - 图	(359)
其他控制图	(360)
补充习题答案	(373)
附录	(388)
附录 I 标准正态分布的分布密度值	(388)
附录 II 标准正态分布的随机变量落在 0 到 z 区间上的概率值	(389)
附录 III t 分布的下侧分位数	(390)
附录 IV χ^2 分布的下侧分位数	(391)
附录 V F 分布的 95% 的下侧分位数	(392)
附录 VI F 分布的 99% 的下侧分位数	(393)
附录 VII 常用对数表	(394)
附录 VIII $e^{-\lambda}$ 值 ($0 < \lambda < 1$)	(397)
附录 IX 随机数表	(398)

第一章 变量和图形

统计学

统计学是一门关于用科学方法收集、整理、汇总、描述和分析数据资料,并在此基础上进行推断和决策的科学.

狭义地说,统计这个术语被用来统指数据或从数据中得到的一些数字,比如平均数.因此,我们常提到职业统计,事故统计等.

总体和样本,归纳统计学和描述性统计学

在收集一组反映人或物的特征的数据时,比如一所大学学生的身高和体重,一个工厂某天生产的螺栓的次品和正品数,观察整组是不可能也是不切实际的,特别是当整组容量很大的时候.常用的方法是观测这个组中的一个部分——**样本**,而不是观测整个组——**总体**.

一个总体可以由有限个元素或无限个元素组成.例如,将一个工厂某天生产的螺栓视为一个总体,那么它是有限的;而将一枚硬币连续抛掷得到的所有结果(正面、反面)视为一个总体,那么它是无限的.

如果样本能很好地反映总体特性,那么就可以通过对样本的分析来给总体下结论,在这种情况下进行的统计工作称为**归纳统计学**或**统计推断**.由于这样的推断不能绝对肯定,因此在下结论时常常用到**概率**这一概念.如果仅仅只是描述和分析特定的对象而不下结论或对较大的群体不进行推断,在这种情况下进行的统计工作称为**描述性统计学**或**演绎统计学**.

在学习统计学之前,我们先来复习一些重要的数学概念.

离散变量和连续变量

一个**变量**用一个符号表示,如 X, Y, H, x 或 B .变量的**值域**是指变量的一切可能取值的集合.如果一个变量仅能取一个值,那么这个变量称为**常量**.

一个理论上可以取 2 个给定值之间任意值的变量称为**连续变量**.反之,则称为**离散变量**.

例 1 一个家庭中儿童人数 N 可取 $0, 1, 2, 3, \dots$ 中任意值,但 N 不能是 2.5 或 3.842,因此 N 是一个离散变量.

例 2 一个人的身高 H ,根据测量的精度,可以是 62 英寸¹⁾,63.8 英寸或 65.8341 英寸, H 是一个连续变量.

可以用离散变量或连续变量来描述的数据分别被称为**离散数据**或**连续数据**.例如,每 1000 个家庭中儿童数量就是一组离散数据,而 100 个大学生的身高就是一组连续数据.通常**测量**产生连续数据,而**枚举**或**计数**产生离散数据.

有时我们还需将变量概念扩展到非数字化实体.例如,彩虹的颜色 C 就是一个变量,它可取“值”为红、橙、黄、绿、蓝、青和紫.通常可以用数值量化这样的变量,比如用 1 表示红色,用 2 表示橙色等等.

数据舍入

由于 72.8 到 73 的距离较它到 72 的距离近,这是因此舍入 72.8 的结果为 73.同样,72.8146 舍入至最近的百分位(或小数点后 2 位)即是 72.81,这是因为 72.8146 距 72.81 较

1) 1 英寸 = 0.0254 米.

72.82 近些.

在把 72.465 舍入至最近的百分位时,我们发现 72.465 距 72.48 和距 72.47 恰好一样远. 在实践中对于这样的情况,常常是把 5 舍入后得到偶数. 因此 72.465 舍入至 72.46, 183.575 舍入至 183.58, 116 500 000 舍入至最近的百万位就是 116 000 000. 这样做的好处就是在大量的运算中,可以使累积舍入误差达到最小(见习题 1.4).

科学记数

在书写数字时,特别是碰到那些小数点前后有许多零的数,我们用科学记数法就方便多了.

例 3 $10^1 = 10$, $10^2 = 10 \times 10 = 100$, $10^5 = 10 \times 10 \times 10 \times 10 \times 10 = 100\,000$,
 $10^8 = 100\,000\,000$.

例 4 $10^0 = 1$; $10^{-1} = .1$ 或 0.1 ; $10^{-2} = .01$ 或 0.01 ; $10^{-5} = .00001$ 或 0.00001 .

例 5 $864\,000\,000 = 8.64 \times 10^8$, $0.00003416 = 3.416 \times 10^{-5}$.

例如,一个数乘以 10^8 ,相当于把小数点向右移动 8 个位置,而一个数乘以 10^{-6} 相当于把小数点向左移动 6 个位置.

我们通常写 0.1253 而不写 .1253 是为了强调小数点前的数字没有被疏忽,而在不产生混淆的情况下,如在表格里,小数点前的 0 可以省略.

通常我们用圆括号或点来表示 2 个或更多的数相乘. 例如: $(5)(3) = 5 \cdot 3 = 5 \times 3 = 15$, $(10)(10)(10) = 10 \cdot 10 \cdot 10 = 10 \times 10 \times 10 = 1000$. 当字母用来表示数字时,圆括号或点可以省略,比如: $ab = (a)(b) = a \cdot b = a \times b$.

科学记数法常常用在计算当中,特别是在对小数点的定位中,我们可以采用如下规则:

$$10^p \cdot 10^q = 10^{p+q}, \quad \frac{10^p}{10^q} = 10^{p-q}$$

其中 p 和 q 是任意数字.

在 10^p 中, p 称为指数,而 10 称为底数.

例 6 $10^3 \times 10^2 = 1000 \times 100 = 100\,000 = 10^5$, 即 10^{3+2} ;

$\frac{10^6}{10^4} = \frac{1\,000\,000}{10\,000} = 100 = 10^2$, 即 10^{6-4} .

例 7 $4\,000\,000 \times 0.0000000002 = (4 \times 10^6) \times (2 \times 10^{-10}) = 4 \times 2 \times 10^6 \times 10^{-10}$
 $= 8 \times 10^{6-10} = 8 \times 10^{-4} = 0.0008$

例 8 $\frac{0.006 \times 80\,000}{0.04} = \frac{6 \times 10^{-3} \times 8 \times 10^4}{4 \times 10^{-2}} = \frac{48 \times 10^1}{4 \times 10^{-2}} = \frac{48}{4} \times 10^{1-(-2)}$
 $= 12 \times 10^3 = 12\,000$

有效数字

如果一个人的身高被精确地记录为 65.4 英寸,这就意味着这个人的真实身高介于 65.35 英寸和 65.45 英寸之间. 这些精确的数字,除去那些小数点定位所需的零,被称为数的有效数字.

例 9 65.4 有 3 个有效数字.

例 10 4.5300 有 5 个有效数字.

例 11 $.0018 = 0.0018 = 1.8 \times 10^{-3}$ 有 2 个有效数字.

例 12 $.001800 = 0.001800 = 1.800 \times 10^{-3}$ 有 4 个有效数字.

枚举(或计数)而非测量得到的数是准确的,并且有无限多个有效数字. 然而在某些情况下,如果没有更多信息很难决定哪些数字是有效的. 例如,数字 186 000 000 可能有 3, 4, ..., 9 个有效数字. 如果认为它有 5 个有效数字,那么把它记为 186.00×10^6 或 1.8600×10^8 会更妥

当些.

数值计算

在进行包括乘、除以及开方运算时,最终结果的有效数字个数不可能比运算数中有效数字个数最少的数的有效数字个数多(见习题 1.9).

例 13 $73.24 \times 4.52 = 331.$

例 14 $1.648/0.023 = 72.$

例 15 $\sqrt{38.7} = 6.22.$

例 16 $8.416 \times 50 = 420.8$ (如果 50 是准确的).

在进行数的加、减运算时,最终结果的小数点以后的有效数字个数不可能比运算数字小数点以后的最少有效数字个数多(见习题 1.10).

例 17 $3.16 + 2.7 = 5.9.$

例 18 $83.42 - 72 = 11.$

例 19 $47.816 - 25 = 22.816$ (如果 25 是准确的).

以上对加法、减法的法则可以推广(见习题 1.11).

函数

若对于变量 X 的每个值,变量 Y 都有一个或更多的值与之对应,则称 Y 是一个关于 X 的函数,并记为 $Y = F(X)$ (读作“ Y 等于 $F(X)$ ”). F 也可用其他字母(G, φ 等)来代替.

变量 X 称为自变量, Y 称为因变量.

若对于 X 的每一个值, Y 有且仅有一个值与之对应,则称 Y 为 X 的单值函数;否则,称 Y 为 X 的多值函数.

例 20 美国的总人口数 P 是时间 t 的函数,记作 $P = F(t)$.

例 21 竖直悬挂的弹簧受到的拉力 S 是挂在它末端的重量 W 的函数,记作: $S = G(W)$.

变量间的函数关系常常用表来描述.然而,它也可以用关于变量的方程来表示,如 $Y = 2X - 3$,在这个方程中,根据不同的 X 的值,就可以得到相应的 Y 的值.

若 $Y = F(X)$,则 $F(3)$ 表示“当 $X = 3$ 时, Y 的值”; $F(10)$ 表示“当 $X = 10$ 时, Y 的值”等等.因此如果 $Y = F(X) = X^2$,那么 $F(3) = 3^2 = 9$,就是当 $X = 3$ 时 Y 的值.

函数的概念可以推广到 2 个或更多个变量(见习题 1.17).

直角坐标

考虑两条相互垂直的带有适当刻度的直线 $X'OX$ 和 $Y'OY$,分别称为 X 轴和 Y 轴(见图

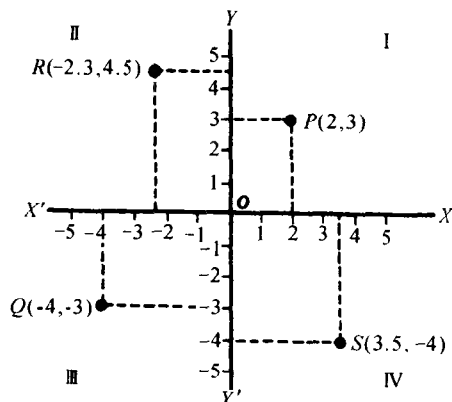


图 1-1

1-1). 整个平面被这两条直线分割成 4 个部分, 分别用 I, II, III 和 IV 来表示, 称为第一、第二、第三和第四**象限**, 整个平面称为 **XY 平面**.

O 点称为**原点**或**零点**. 过任意给定的一点 P, 向 X 轴和 Y 轴分别作垂线. 垂线与 X 轴的交点值 X 和与 Y 轴的交点值 Y 称为 P 点的**直角坐标**, 或简称为**坐标**, 记为 (X, Y). X 称为点的**横坐标**, Y 称为点的**纵坐标**. 在图 1-1 中, P 点的横坐标是 2, 纵坐标是 3, P 点的坐标记为 (2, 3).

相反地, 给定了点的坐标, 我们就可以对这个点定位或**作图**. 在图 1-1 中, Q, R 和 S 分别表示坐标为 $(-4, -3)$, $(-2.3, 4.5)$ 和 $(3.5, -4)$ 的点.

过 O 点作 Z 轴垂直于 XY 平面, 我们可以很容易地推广以上的概念到三维空间. 在这种情况下, 点 P 的坐标应为 (X, Y, Z).

图形

变量间的关系也可用**图形**来表示. 根据数据的性质和应用图形的目的, 有许多种类的图形已在统计学中得到使用. 这其中有**条形图**、**圆形图**、**像形图**等. 这些图形也常称为**图表**, 例如条形图表, 圆形图表等(见习题 1.23, 1.24, 1.26 和 1.27).

方程

方程是形如 $A = B$ 的表达式, 其中 A 称为**方程的左端(边)**, B 称为**方程的右端(边)**. 只要我们对方程的两边进行**相同运算**, 我们就能得到**等价方程**. 因此我们对方程的两边同时加上、减去、乘以或除以同一个值, 我们就能得到一个等价方程. 要注意的是**除以 0 是不允许的**.

例 22 对于方程 $2X + 3 = 9$, 从两边同时减去 3, $2X + 3 - 3 = 9 - 3$ 得 $2X = 6$. 两边同时除以 2, $2X/2 = 6/2$, 得 $X = 3$. 这个值就是给定方程的**解**. 用 3 来代替 X, 得 $2 \times 3 + 3 = 9$, 即 $9 = 9$, 这是一个**恒等式**. 求方程解的过程称为**解方程**.

上述方法可以解含有 2 个未知数的 2 个方程, 含有 3 个未知数的 3 个方程等等. 这样的一组方程称为**联立方程**(见习题 1.30).

不等式

符号“ $<$ ”和“ $>$ ”分别表示“小于”和“大于”. 符号“ \leq ”和“ \geq ”则分别表示“小于等于”和“大于等于”. 众所周知, 这些都是**不等号**.

例 23 $3 < 5$ 读作“3 小于 5”.

例 24 $5 > 3$ 读作“5 大于 3”.

例 25 $X < 8$ 读作“X 小于 8”.

例 26 $X \geq 10$ 读作“X 大于等于 10”.

例 27 $4 < Y \leq 6$ 读作“4 小于 Y, Y 小于等于 6”或“Y 在 4 和 6 之间, 不包括 4, 而包括 6”, 或“Y 大于 4 且小于等于 6”.

有不等号连接的关系式称为**不等式**. 与等式的两边一样, 我们也称不等式的两边为**不等式的端**, 因此在不等式 $4 < Y \leq 6$ 中, 4, Y 和 6 是不等式的各端.

在下列情况下, 不等式依然成立:

1. 不等式各端同时加上或减去相同的数.

例 28 由于 $15 > 12$, 因此 $15 + 3 > 12 + 3$ (即 $18 > 15$) 和 $15 - 3 > 12 - 3$ (即 $12 > 9$).

2. 不等号各端同时乘以或除以相同的**正数**.

例 29 由于 $15 > 12$, 因此 $15 \times 3 > 12 \times 3$ (即 $45 > 36$) 和 $15/3 > 12/3$ (即 $5 > 4$).

3. 不等式各端同时乘以或除以相同的**负数**, 不等号的方向要改变.

例 30 由于 $15 > 12$, 因此 $15 \times (-3) < 12 \times (-3)$ (即 $-45 < -36$) 和 $15/(-3) > 12/(-3)$ (即 $-5 < -4$).

对数

每个正数 N 都可以表示为 10 的幂的形式,也就是说,我们总可以找到 p ,使得 $N = 10^p$, p 称为以 10 为底的 N 的对数或 N 的常用对数,记为 $p = \log_{10} N$,或简记为 $p = \log N$.例如,由于 $1000 = 10^3$,因此 $\log 1000 = 3$.同样地,由于 $0.01 = 10^{-2}$,因此 $\log 0.01 = -2$.

当 N 是 1 到 10(即 10^0 到 10^1)之间的数时, $p = \log N$ 是 0 到 1 之间的数,并可以在附录 VII 中查找.

例 31 在附录 VII 中查找 $\log 2.36$,顺着标有 N 的左列向下看,直到我们找到第一个 2 位数 23,然后我们向右直到看到标有 6 的列,就找到了表值 3729.因此 $\log 2.36 = 0.3729$ (即 $2.36 = 10^{0.3729}$).

所有正数的对数都可以从 1 到 10 的对数中得到.

例 32 在例 31 中, $2.36 = 10^{0.3729}$,用 10 连续地去乘,得到 $23.6 = 10^{1.3729}$, $236 = 10^{2.3729}$, $2360 = 10^{3.3729}$ 等等.因此 $\log 2.36 = 0.3729$, $\log 23.6 = 1.3729$, $\log 236 = 2.3729$, $\log 2360 = 3.3729$.

例 33 由于 $2.36 = 10^{0.3729}$,我们连续地用 10 去除得到 $0.236 = 10^{0.3729-1} = 10^{-0.6271}$, $0.0236 = 10^{0.3729-2} = 10^{-1.6271}$,以此类推.

通常我们把 $0.3729 - 1$ 写为 $9.3729 - 10$ 或 $\bar{1}.3729$; $0.3729 - 2$ 写为 $8.3729 - 10$ 或 $\bar{2}.3729$,以此类推.有了这些记号,我们有

$$\log 0.236 = 9.3729 - 10 = \bar{1}.3729 = -0.6271$$

$$\log 0.0236 = 8.3729 - 10 = \bar{2}.3729 = -1.6271$$

以此类推.

小数部分 .3729 在这些算法中称为**对数的尾数**.在尾数小数点前的部分,[即 1, 2, 3, $\bar{1}$, $\bar{2}$ (或 $9 - 10$, $8 - 10$)]称为**首数**.

以下法则很容易证明:

1. 大于 1 的数,它的首数是非负的,并且首数比小数点前的数的位数小.

例 34 2360, 236, 23.6 和 2.36 的对数的首数分别为 3, 2, 1 和 0,而得到的对数分别为 3.3729, 2.3729, 1.3729 和 0.3729.

2. 小于 1 的数,它的首数是负的,并且首数的绝对值比紧跟着小数点后的零的个数大.

例 35 0.236, 0.0236 和 0.00236 的首数分别为 -1 , -2 , -3 ,而求得的对数分别为 $\bar{1}.3729$, $\bar{2}.3729$ 和 $\bar{3}.3729$ 或 $9.3729 - 10$, $8.3729 - 10$ 和 $7.3729 - 10$.

如果要算 4 个数字数的对数,比如 2.364 和 758.2,可以用**插值法**求解(见习题 1.36).

反对数

在指数形式 $2.36 = 10^{0.3729}$ 中, 2.36 称为 0.3729 的反对数.2.36 的对数是 0.3729,接下来就可以得到

$$\text{antilog } 1.3729 = 23.6, \text{ antilog } 2.3729 = 236, \text{ antilog } 3.3729 = 2360$$

$$\text{antilog } 9.3729 - 10 = \text{antilog } \bar{1}.3729 = 0.236$$

$$\text{antilog } 8.3729 - 10 = \text{antilog } \bar{2}.3729 = 0.0236$$

任何数的反对数都可通过附录 VII 查找.

例 36 求 $\text{antilog } 8.6284 - 10$,在表中寻找尾数 0.6284.它所处的行标号为 42,而列标号为 5,所以所求的数为 425.由于首数为 $8 - 10$,因此所求的结果是 0.0425.同样地, $\text{antilog } 3.6284 = 4250$, $\text{antilog } 5.6284 = 425\ 000$.

如果在附录 VII 中找不到某个尾数,那么可以使用插值法(见习题 1.37).

对数计算

对数计算遵循下列法则：

$$\log MN = \log M + \log N$$

$$\log \frac{M}{N} = \log M - \log N$$

$$\log M^p = p \log M$$

根据这些结论, $\log \frac{A^p B^q C^r}{D^s E^t} = p \log A + q \log B + r \log C - s \log D - t \log E$ (见习题 1.38 ~ 1.45).

习题及解答

变量

1.1 说出下列哪些是离散数据, 哪些是连续数据:

- (a) 股票市场每天抛售的股票数;
- (b) 气象局每隔半个小时记录的气温;
- (c) 某公司生产的电视显像管的寿命;
- (d) 大学教授的年收入;
- (e) 某工厂生产的 1000 个螺栓的长度.

解 (a) 离散数据; (b) 连续数据; (c) 连续数据; (d) 离散数据; (e) 连续数据.

1.2 给出下列变量的值域, 并说出哪些是连续变量, 哪些是离散变量:

- (a) 一台洗衣机里水的加仑数 G ;
- (b) 一书架上书的数目 B ;
- (c) 掷一对骰子所得的点数之和 S ;
- (d) 一个球的直径 D ;
- (e) 欧洲的国家 C .

解 (a) 值域: 从 0 到洗衣机容量之间的任意值. G 是连续变量.

(b) 值域: 0, 1, 2, 3, ... 直到书架上最多能放置的书的数目. B 是离散变量.

(c) 值域: 一个骰子上的点数可以是 1, 2, 3, 4, 5 或 6. 因此掷一对骰子所得的点数之和 S 可以为 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 或 12. S 是离散变量.

(d) 值域: 如果我们把一点视为直径是 0 的球, 那么 D 的值域为大于等于 0 的一切实数. D 是连续变量.

(e) 值域: C 可以为英国、法国、德国等, 可分别用 1, 2, 3 等来表示. C 是离散变量.

数据舍入

1.3 按要求舍入下列数据:

- (a) 48.6 舍入至最近的整数; (f) 143.95 舍入至最近的十分位数;
- (b) 136.5 舍入至最近的整数; (g) 368 舍入至最近的百位数;
- (c) 2.484 舍入至最近的百分位数; (h) 24 448 舍入至最近的千位数;
- (d) 0.0435 舍入至最近的千分位数; (i) 5.56500 舍入至最近的百分位数;
- (e) 4.50001 舍入至最近的整数; (j) 5.56501 舍入至最近的百分位数.

解 (a) 49; (b) 136; (c) 2.48; (d) 0.044; (e) 5; (f) 144.0; (g) 400; (h) 24 000; (i) 5.56; (j) 5.57.

1.4 按要求对 4.35, 8.65, 2.95, 12.45, 6.65, 7.55 和 9.75 做加法. (a) 直接加; (b) 根据“偶数”原则舍入至最近的十分位数; (c) 遇 5 向前进.

解 (a)	4.35	(b)	4.4	(c)	4.4
	8.65		8.6		8.7
	2.95		3.0		3.0
	12.45		12.4		12.5
	6.65		6.6		6.7
	7.55		7.6		7.6
	9.75		9.8		9.8

总和 52.35

总和 52.4

总和 52.7

注意,过程(b)的累积舍入误差最小,因此过程(b)优于过程(c).

科学记数和有效数字

1.5 下列数字若不用 10 的幂次方表示各为多少:

- (a) 4.823×10^7 ; (c) 3.80×10^{-4} ; (e) 300×10^8 ;
 (b) 8.4×10^{-6} ; (d) 1.86×10^5 ; (f) $70\,000 \times 10^{-10}$.

解 (a)把小数点向右移 7 个位置,得到 48 230 000;

(b)把小数点向左移 6 个位置,得到 0.0000084;

(c)0.000380;(d)186 000;

(e)30 000 000 000;(f)0.0000070000.

1.6 假定下列数是精确记录的,每个数有几个有效数字?

- (a)149.8 英寸; (d)0.00280 米; (g)9 所房子;
 (b)149.80 英寸; (e)1.00280 米; (h) 4.0×10^3 磅;
 (c)0.0028 米; (f)9 克; (i) 7.58400×10^{-5} 达因.

解 (a)4;(b)5;(c)2;(d)3;(e)6;(f)1;(g)无限;(h)2;(i)6.

1.7 假定以下测量值是精确记录的,每个测量值的最大误差是多少?

- (a)73.854 英寸;(b)0.09800 立方英尺;(c) 3.867×10^8 公里.

解 (a)此测量值可以从 73.8535 到 73.8545 英寸取值,因此最大误差是 0.0005 英寸.共有 5 个有效数字.

(b)此测量值可以从 0.097995 到 0.098005 立方英尺取值,因此最大误差是 0.000005 立方英尺.共有 4 个有效数字.

(c)准确的公里数大于 3.8665×10^8 而小于 3.8675×10^8 ,因此最大误差是 0.0005×10^8 或 50 000 公里.共有 4 个有效数字.

1.8 用科学记数法书写下列数字(除非另外说明,否则认为所有数据是精确记录的).

- (a)24 380 000(4 个有效数字);(b)0.000009851;(c)7 300 000 000(5 个有效数字);
 (d)0.00018400.

解 (a) 2.438×10^7 ; (b) 9.851×10^{-6} ; (c) 7.3000×10^9 ; (d) 1.8400×10^{-4} .

数值计算

1.9 证明如果假设 5.74 和 3.8 分别有 3 个和 2 个有效数字,那么它们的乘积不能精确到多于两个有效数字.

解 解法一 $5.74 \times 3.8 = 21.812$,但并不是积的所有数字都是有效数字.由于 5.74 代表了 5.735 到 5.745 之间的任何数.而 3.8 代表了 3.75 到 3.85 之间的任何数,因此 5.74 和 3.8 的积的最小可能值为 $5.735 \times 3.75 = 21.50625$,最大可能值为 $5.745 \times 3.85 = 22.11825$.

由于积的可能范围从 21.50625 到 22.11825,因此最多只有积的前两个数是有效的,结果应为 22.数字 22 代表了从 21.5 到 22.5 之间的任何数.

解法二 用斜体字表示疑问数字,乘法可如下进行:

$$\begin{array}{r} 5.74 \\ \times 3.8 \\ \hline 4592 \\ 1722 \\ \hline 21.812 \end{array}$$

在答案里不应有多余一个的怀疑数字,因此取2个有效数字22.注意,在运算中,没有必要运算比数据中小数点后最小有效数字个数更多的有效数字,因此如果5.74舍入到5.7,积为 $5.7 \times 3.8 = 21.66 = 22$,有两个有效数字,与上述结果一样.

如果不用电脑计算,那么在计算中不需要保留比小数点后最小有效数字个数多于一个或两个的有效数字,并把最终答案舍入至适当的数.用计算机计算时,它能提供许多位数字,我们必须注意并不是所有数字都是有效的.

1.10 假设所有数字是有效的,对4.19355, 15.28, 5.9561, 12.3和8.472做加法.

解 (a) 在做加法运算中,用斜体字表示疑问数字.不超过一个疑问数字的最终答案为46.2.

$$\begin{array}{r} \text{(a)} \quad \begin{array}{r} 4.19355 \\ 15.28 \\ 5.9561 \\ 12.3 \\ 8.472 \\ \hline 46.20165 \end{array} \quad \text{(b)} \quad \begin{array}{r} 4.19 \\ 15.28 \\ 5.96 \\ 12.3 \\ 8.47 \\ \hline 46.20 \end{array} \end{array}$$

(b)在加法运算中保留比小数点后最小有效数字个数多一位的有效数字,这样就可以减少一些计算量.最后结果舍入至46.2,与(a)中结果一致.(b)的运算较简便.

1.11 计算 $475\,000\,000 + 12\,684\,000 - 1\,372\,410$,假设这些数分别有3个、5个和7个有效数字.

解 在(a)的运算中,所有的数字被保留,最后结果被舍入.在(b)的运算中,使用类似于习题1.10(b)的方法.在两种方法中,用斜体字表示疑问数字.

$$\begin{array}{r} \text{(a)} \quad \begin{array}{r} 475000000 \\ + 12684000 \\ \hline 487684000 \end{array} \quad \begin{array}{r} 487684000 \\ - 1372410 \\ \hline 486311590 \end{array} \\ \text{(b)} \quad \begin{array}{r} 475000000 \\ + 12700000 \\ \hline 487700000 \end{array} \quad \begin{array}{r} 487700000 \\ - 1400000 \\ \hline 486300000 \end{array} \end{array}$$

最后结果舍入至486 000 000;或为了突出它有了个有效数字,可写为 4.86×10^8 .

1.12 完成下列运算.

$$\begin{array}{ll} \text{(a)} 48.0 \times 943; & \text{(e)} \frac{(1.47562 - 1.47322) \times 4895.36}{0.000159180}; \\ \text{(b)} 8.35/98; & \text{(f)} \text{若分母5和6是准确的, } \frac{(4.38)^2}{5} + \frac{(5.482)^2}{6}; \\ \text{(c)} 28 \times 4193 \times 182; & \text{(g)} 3.1416 \sqrt{71.35}; \\ \text{(d)} \frac{526.7 \times 0.001280}{0.000034921}; & \text{(h)} \sqrt{128.5 - 89.24}. \end{array}$$

解 (a) $48.0 \times 943 = 45\,300$

(b) $8.35/98 = 0.085$

$$\begin{aligned} \text{(c)} \quad 28 \times 4193 \times 182 &= 2.8 \times 10^1 \times 4.193 \times 10^3 \times 1.82 \times 10^2 \\ &= 2.8 \times 4.193 \times 1.82 \times 10^{1+3+2} = 21 \times 10^6 \\ &= 2.1 \times 10^7 \end{aligned}$$

$$\begin{aligned}
 \text{(d)} \quad \frac{526.7 \times 0.001280}{0.000034921} &= \frac{5.267 \times 10^2 \times 1.280 \times 10^{-3}}{3.4921 \times 10^{-5}} \\
 &= \frac{5.267 \times 1.280}{3.4921} \times \frac{10^2 \times 10^{-3}}{10^{-5}} \\
 &= 1.931 \times \frac{10^{2-3}}{10^{-5}} = 1.931 \times \frac{10^{-1}}{10^{-5}} = 1.931 \times 10^{-1+5} \\
 &= 1.931 \times 10^4
 \end{aligned}$$

$$\begin{aligned}
 \text{(e)} \quad \frac{1.47562 - 1.47322 \times 4895.36}{0.000159180} &= \frac{0.00240 \times 4895.36}{0.000159180} \\
 &= \frac{2.40 \times 10^{-3} \times 4.89536 \times 10^3}{1.59180 \times 10^{-4}} \\
 &= \frac{2.40 \times 4.89536}{1.59180} \times \frac{(10^{-3} \times 10^3)}{10^{-4}} \\
 &= 7.38 \times \frac{10^0}{10^{-4}} \\
 &= 7.38 \times 10^4
 \end{aligned}$$

写成 7.38×10^4 表示 3 个有效数字. 注意, 尽管一开始所有的数字有 6 个有效数字, 但从 1.47562 减去 1.47322 时, 它们中的一些就消失了.

$$\text{(f)} \quad \text{若分母 5 和 6 是准确的, } \frac{(4.38)^2}{5} + \frac{(5.482)^2}{6} = 3.84 + 5.009 = 8.85.$$

$$\text{(g)} \quad 3.1416 \sqrt{71.35} = 3.1416 \times 8.447 = 26.54$$

$$\text{(h)} \quad \sqrt{128.5 - 89.24} = \sqrt{39.3} = 6.27$$

1.13 当 $X = 3, Y = -5, A = 4, B = -7$ 时, 计算下列各式的值:

$$\text{(a)} 2X - 3Y; \quad \text{(e)} 2(X + 3Y) - 4(3X - 2Y);$$

$$\text{(b)} 4Y - 8X + 28; \quad \text{(f)} \frac{X^2 - Y^2}{A^2 - B^2 + 1};$$

$$\text{(c)} \frac{AX + BY}{BX - AY}; \quad \text{(g)} \sqrt{2X^2 - Y^2 - 3A^2 + 4B^2 + 3};$$

$$\text{(d)} X^2 - 3XY - 2Y^2; \quad \text{(h)} \sqrt{\frac{6A^2}{X} + \frac{2B^2}{Y}}.$$

解 $\text{(a)} 2X - 3Y = 2 \times 3 - 3 \times (-5) = 6 + 15 = 21$

$$\text{(b)} 4Y - 8X + 28 = 4 \times (-5) - 8 \times 3 + 28 = -20 - 24 + 28 = -16$$

$$\text{(c)} \frac{AX + BY}{BX - AY} = \frac{4 \times 3 + (-7) \times (-5)}{(-7) \times 3 - 4 \times (-5)} = \frac{12 + 35}{-21 + 20} = \frac{47}{-1} = -47$$

$$\text{(d)} X^2 - 3XY - 2Y^2 = 3^2 - 3 \times 3 \times (-5) - 2 \times (-5)^2 = 9 + 45 - 50 = 4$$

$$\begin{aligned}
 \text{(e)} \quad 2(X + 3Y) - 4(3X - 2Y) &= 2[3 + 3 \times (-5)] - 4[3 \times 3 - 2 \times (-5)] \\
 &= 2 \times (3 - 15) - 4 \times (9 + 10) = 2 \times (-12) - 4 \times 19 \\
 &= -24 - 76 = -100
 \end{aligned}$$

另解 $2(X + 3Y) - 4(3X - 2Y) = 2X + 6Y - 12X + 8Y = -10X + 14Y$
 $= -10(3) + 14(-5) = -30 - 70 = -100$

$$\text{(f)} \frac{X^2 - Y^2}{A^2 - B^2 + 1} = \frac{(3)^2 - (-5)^2}{(4)^2 - (-7)^2 + 1} = \frac{9 - 25}{16 - 49 + 1} = \frac{-16}{-32} = \frac{1}{2} = 0.5$$

$$\begin{aligned}
 \text{(g)} \quad \sqrt{2X^2 - Y^2 - 3A^2 + 4B^2 + 3} &= \sqrt{2(3)^2 - (-5)^2 - 3(4)^2 + 4(-7)^2 + 3} \\
 &= \sqrt{18 - 25 - 48 + 196 + 3} = \sqrt{144} = 12
 \end{aligned}$$

$$\text{(h)} \sqrt{\frac{6A^2}{X} + \frac{2B^2}{Y}} = \sqrt{\frac{6 \times 4^2}{3} + \frac{2 \times (-7)^2}{-5}} = \sqrt{\frac{96}{3} + \frac{98}{-5}} = \sqrt{12.4} \approx 3.52$$

函数和图形

1.14 表 1.1 记录的是 PQR 农场 1987~1997 年小麦和玉米的产量, 计量单位蒲式耳(bu). 根据这张表判断(a)小麦产量最低的年份;(b)玉米产量最高的年份;(c)小麦减产最多的年份;(d)上述年份中哪一年玉米减产而小麦增产;(e)小麦产量相等的年份;(f)小麦和

玉米总产量达到最大的年份.

表 1.1

年份	小麦年产量 (舍入至最近的 5 蒲式耳)	玉米年产量 (舍入至最近的 5 蒲式耳)
1987	200	75
1988	185	90
1989	225	100
1990	250	85
1991	240	80
1992	195	100
1993	210	110
1994	225	105
1995	250	95
1996	230	110
1997	235	100

解 (a) 1988 年; (b) 1993 和 1996 年; (c) 1992 年; (d) 1990, 1994, 1995 和 1997 年; (e) 1989 和 1994 年, 1990 和 1995 年; (f) 1995 年.

1.15 W 和 C 分别表示习题 1.14 中 PQR 农场的小麦和玉米的产量. 显然 W 和 C 都是关于年份 t 的函数, 用 $W = F(t)$ 和 $C = G(t)$ 表示.

- (a) 当 $t = 1993$ 时, 分别求 W ;
- (b) 当 $t = 1990$ 和 1996 时, 求 C ;
- (c) 当 $W = 225$ 时, 求 t ;
- (d) 求 $F(1991)$;
- (e) 求 $G(1995)$;
- (f) 当 $W = 210$ 时, 求 C ;
- (g) 变量 t 的定义域是什么?
- (h) W 是 t 的单值函数吗?
- (i) t 是 W 的函数吗? 若是, 是单值函数吗?
- (j) C 是 W 的函数吗?
- (k) t 和 W 哪个是自变量?

解 (a) 210; (b) 分别为 85 和 110; (c) 1989 和 1994; (d) 240; (e) 95; (f) 110; (g) 1987~1997; (h) 是. 对于定义域里的每个 t , W 有且只有一个值与之对应. (i) 是. 因为对 W 的每个值, t 都有多于一个的值与之对应(例如, $W = 225$ 时, $t = 1989$ 和 $t = 1994$), 这个函数是多值的, 可写为 $t = H(W)$. (j) 是. 因为对 W 的每个值, C 都有一个或更多的值与之对应, 如表 1.1 所示. 同理, W 是 C 的函数. (k) 通常我们会认为 W 是由 t 决定的, 而不认为 t 是由 W 所决定的. 因此, t 是自变量, W 是因变量. 然而, 从数学角度来看, 任何一个变量可视为自变量, 另外一个则视为因变量. 可取不同值的变量为自变量, 由它决定的变量为因变量.

1.16 变量 Y 由变量 X 确定, 二者的关系由等式 $Y = 2X - 3$ 决定.

- (a) 当 $X = 3, -2$ 和 1.5 时, 求 Y ;
- (b) 当 $X = -2, -1, 0, 1, 2, 3$ 和 4 时, 列表求 Y ;
- (c) 若 $Y = F(X)$, 求 $F(2.4)$ 和 $F(0.8)$;
- (d) 当 $Y = 15$ 时, 求 X ;
- (e) X 能表示为 Y 的函数吗?
- (f) Y 是 X 的单值函数吗?
- (g) X 是 Y 的单值函数吗?

解 (a) $X = 3, Y = 2X - 3 = 2 \times 3 - 3 = 6 - 3 = 3$,
 $X = -2, Y = 2X - 3 = 2 \times (-2) - 3 = -4 - 3 = -7$,
 $X = 1.5, Y = 2X - 3 = 2 \times 1.5 - 3 = 3 - 3 = 0$.

(b) 根据(a)中的计算, Y 的值列在表 1.2 中. 注意到根据其他的 X 值, 我们还可列出许多表.

表 1.2

X	-2	-1	0	1	2	3	4
Y	-7	-5	-3	-1	1	3	5

(c) $F(2.4) = 2 \times 2.4 - 3 = 4.8 - 3 = 1.8$,

$F(0.8) = 2 \times 0.8 - 3 = 1.6 - 3 = -1.4$.

(d) $Y = 15$ 代入 $Y = 2X - 3$, $15 = 2X - 3$, $2X = 18$, $X = 9$.

(e) 是. 由于 $Y = 2X - 3$, $Y + 3 = 2X$, 即 $X = \frac{1}{2}(Y + 3)$. X 是关于 Y 的显函数.

(f) 是. 对于定义域里的每个 X , Y 有且只有一个值与之对应.

(g) 是. 根据 $X = \frac{1}{2}(Y + 3)$, 对于每个 Y , X 有且只有一个值与之对应.

- 1.17 若 $Z = 16 + 4X - 3Y$, 求相应的 Z 值. (a) $X = 2$, $Y = 5$; (b) $X = -3$, $Y = -7$; (c) $X = -4$, $Y = 2$.

解 (a) $Z = 16 + 4 \times 2 - 3 \times 5 = 16 + 8 - 15 = 9$;

(b) $Z = 16 + 4 \times (-3) - 3 \times (-7) = 16 - 12 + 21 = 25$;

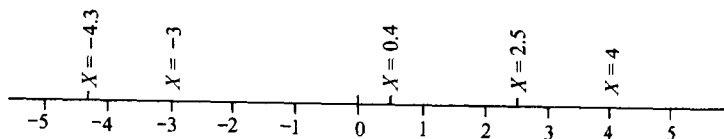
(c) $Z = 16 + 4 \times (-4) - 3 \times 2 = 16 - 16 - 6 = -6$.

给定 X 和 Y , 就有相应的 Z . Z 与 X 和 Y 的联系用 $Z = F(X, Y)$ 来表示 (读作“ Z 是关于 X 和 Y 的函数”). $F(2, 5)$ 表示当 $X = 2$, $Y = 5$ 时, Z 的值, 这个值为 9. 同理, $F(-3, -7) = 25$, $F(-4, 2) = -6$.

X 和 Y 称为自变量, Z 称为因变量.

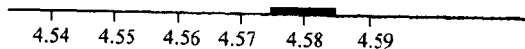
- 1.18 在坐标系统的 X 轴上根据坐标确定点的位置 (a) $X = 4$; (b) $X = -3$; (c) $X = 2.5$; (d) $X = -4.3$; (e) $X = 0.4$. 假设所有数据均是准确的.

解 对 X 的每一个准确值, 在 X 轴上都有且只有一个点与之对应. 用进一步的数学知识还可证明在轴上的任一个点都有且只有一个 X 值与之对应.



因此, 从理论上讲, 有一个点与 $X = 22/7 = 3.142857142857 \dots$ 相对应, 有另一个点与 $X = \pi = 3.14159265358 \dots$ 相对应. 当然在应用中, 由于我们的笔迹有厚度并且它覆盖了无穷多个点, 因此我们不可能找出一个点的确切位置. X 轴本身也有厚度. 所以我们所作的图只是实际数学情况的形式代表.

- 1.19 X 表示一滚珠轴承直径, 单位是厘米 (cm). 如果 $X = 4.58$ 有三个有效数字, 那么如何在 X 轴上表示?



解 4.58 cm 的真实测量值处在 4.575 cm 和 4.585 cm 之间, 用粗线段在上图中表示.

- 1.20 在直角坐标系中找出下列点的位置: (a) $(5, 2)$; (b) $(2, 5)$; (c) $(-3, 1)$; (d) $(1, -3)$; (e) $(3, -4)$; (f) $(-2.5, -4.8)$; (g) $(0, -2.5)$; (h) $(4, 0)$. 假定所有数均是准确的.

解 见图 1-2.

- 1.21 画出方程 $Y = 2X - 3$ 的图.

解 找出 $X = -2, -1, 0, 1, 2, 3$ 和 4, 分别求出 $Y = -7, -5, -3, -1, 1, 3$ 和 5 (见习题 1.16 (b)). 因此图上的点为 $(-2, -7)$, $(-1, -5)$, $(0, -3)$, $(1, -1)$, $(2, 1)$, $(3, 3)$ 和 $(4, 5)$, 把它们标于直角坐标系中, 如图 1-3 所示. 所有这些点以及用其他 X 值算出的点均在一条直线上, 这条直线即为所

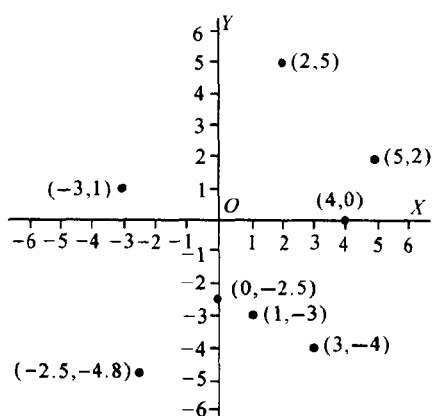


图 1-2

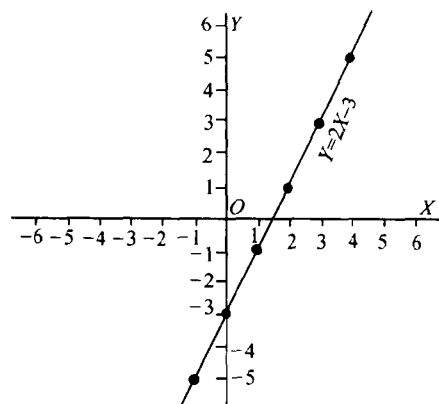


图 1-3

求直线.

因为 $Y = 2X - 3$ 的图像为一条直线, 所以我们称 $F(X) = 2X - 3$ 为**线性函数**. 一般地, $F(X) = aX + b$ (a 和 b 是任意常数) 是线性函数, 它的图像是一条直线.

实际上, 只要两个点就能确定一个线性函数, 因为两点确定一条直线.

1.22 作方程 $Y = X^2 - 2X - 8$ 的图像.

解 根据不同的 X 值, 对应的 Y 值列在表 1.3 中. 例如, 当 $X = -2$ 时, $Y = (-2)^2 - 2 \times (-2) - 8 = 4 + 4 - 8 = 0$. 从表中可以看出, $(-3, 7)$, $(-2, 0)$, $(-1, -5)$, $(0, -8)$, $(1, -9)$, $(2, -8)$, $(3, -5)$, $(4, 0)$ 和 $(5, 7)$ 这些点以及用其他 X 值算出的点均在如图 1-4 所示的曲线中. 这条曲线称为**抛物线**. 函数 $F(x) = X^2 - 2X - 8$ 称为**二次函数**.

表 1.3

X	-3	-2	-1	0	1	2	3	4	5
Y	7	0	-5	-8	-9	-8	-5	0	7

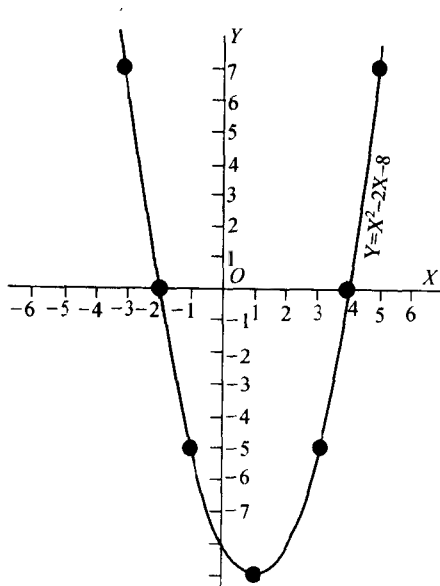


图 1-4

一般地,方程 $Y = a + bX + cX^2$ (其中 a, b 和 c 是常数, 并且 $c \neq 0$) 的图像是抛物线. 如果 $c = 0$, 那么图像是一条直线, 参见习题 1.21.

1.23 表 1.4 给出了从 1990 到 1994 年 HIV 患者出院的人数(以千计). 根据数据作图.

表 1.4

年份	1990	1991	1992	1993	1994
HIV 患者出院人数	146	165	194	225	234

来源: 美国国家健康统计中心, 健康与生死统计.

解 解法一 观察图 1-5. 在此图中, HIV 病人出院的人数是一个因变量, 时间是自变量. 点的坐标按照坐标表示法的习惯确定, 如(1990, 146). 用直线把点连接起来. 此图像称为**线图**.

注意, 轴上的单位长度是不等的. 这是因为 2 个变量代表的是不同的数量. 注意图中的零点在竖轴而不在横轴上. 通常, 零点应尽可能地表示出来, 尤其是在竖轴上. 如果在某些情况下无法表示出零点, 并且这样的忽略会造成读者的误解, 那么最好用某些方法提醒读者注意这里的忽略. 有一些变量是时间的函数, 表示这种函数分布的表或图称为**时间序列**.

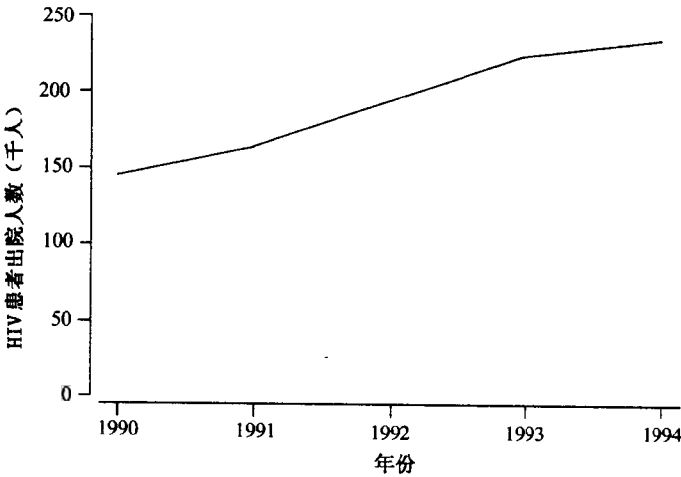


图 1-5 HIV 患者出院人数(来源: 美国国家健康统计中心, 健康与生死统计)

解法二 图 1-6 称为**条形图**. 只要条与条之间不相交, 条带宽可以制成任意方便的尺寸.

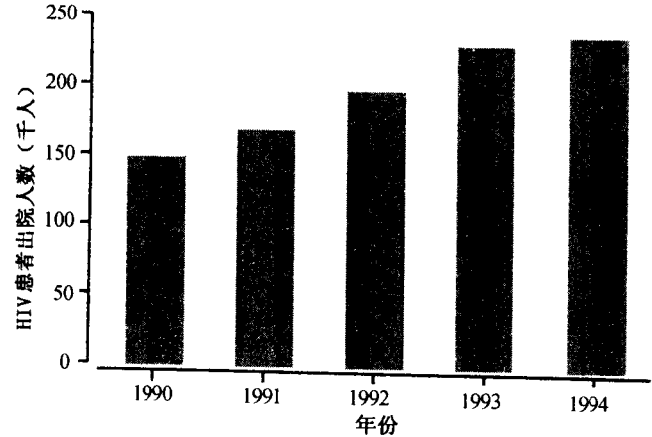


图 1-6 HIV 患者出院人数(来源: 美国国家健康统计中心, 健康与生死统计)

解法三 条带是从横轴上延伸出来而不是从竖轴上延伸出来的条形图,如图 1-7 所示。

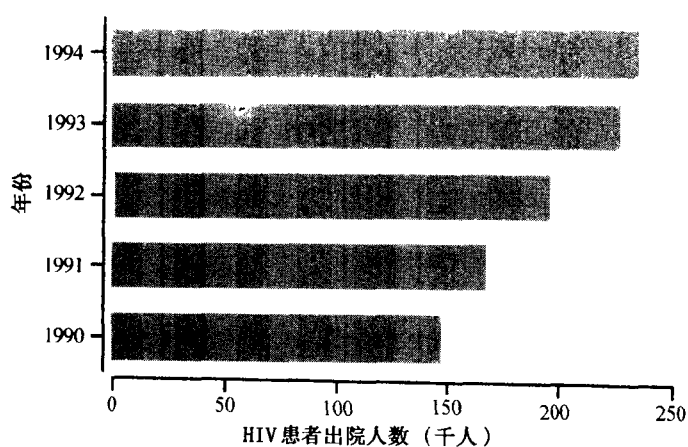


图 1-7 HIV 患者出院人数(来源:美国国家健康统计中心,健康与生死统计)

1.24 根据习题 1.14 的数据,做(a)线图,(b)条形图.

解 (a)线图如图 1-8 所示.

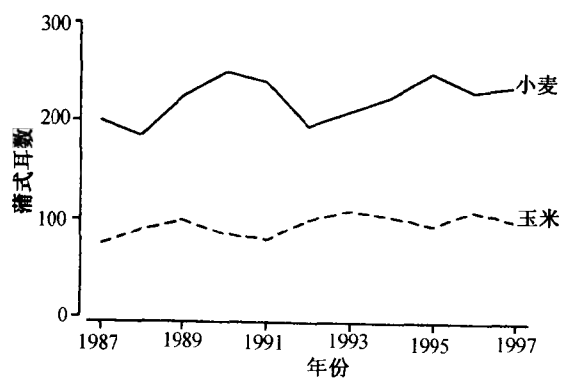


图 1-8

(b)图 1-9 和 1-10 展示了两种类型的条形图.图 1-10 中的图形称为**分支条形图**.

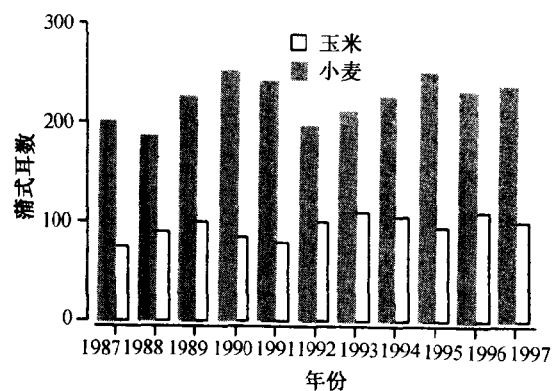


图 1-9

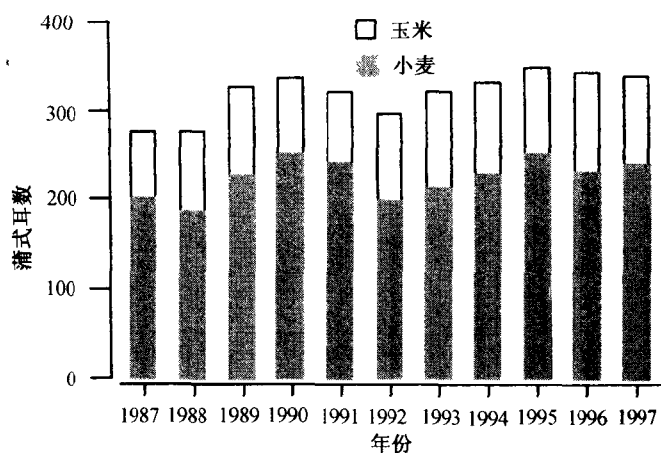


图 1-10

- 1.25 (a)根据习题 1.14 的表 1.1, 写出小麦和玉米的年产量占全年产量的百分比.
(b)根据(a)得到的数据作百分比图.

解 (a)观察 1987 年, 小麦所占百分比为 $200/(200 + 75) = 72.7\%$, 玉米所占的百分比为 $100\% - 72.7\% = 27.3\%$. 百分比如表 1.5 所示.

表 1.5

年份	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
小麦(%)	72.7	67.3	69.2	74.6	75.0	66.1	65.6	68.2	72.5	67.7	70.1
玉米(%)	27.3	32.7	30.8	25.4	25.0	33.9	34.4	31.8	27.5	32.3	29.9

(b)由(a)而得的百分比图如图 1-11 所示, 称为百分数分支图. 也可作类似于图 1-9 的条形图.

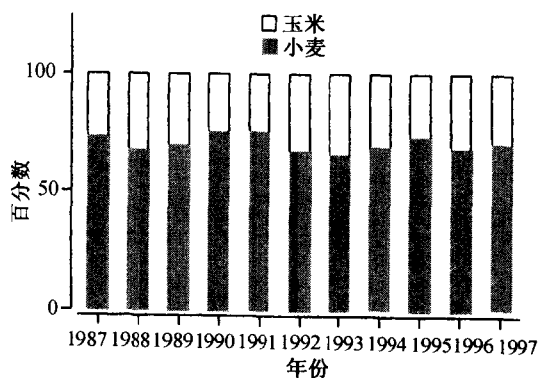


图 1-11

- 1.26 根据表 1.1 中小麦产量作线图.

解 所求直线可通过去掉图 1-8 中较低的那条线图而得. 结果在线图和横轴之间有大片空间被浪费了. 为了避免产生这种情况, 竖轴可以从 150 蒲式耳开始, 而不从 0 蒲式耳开始. 然而, 这也许会给没有注意到零点的读者带来误解. 为了提醒读者, 我们可以如图 1-12 那样作图. 另一种提醒大家注意零点被忽视的方法是在一条轴上用锯齿形线表示, 参见图 1-13.

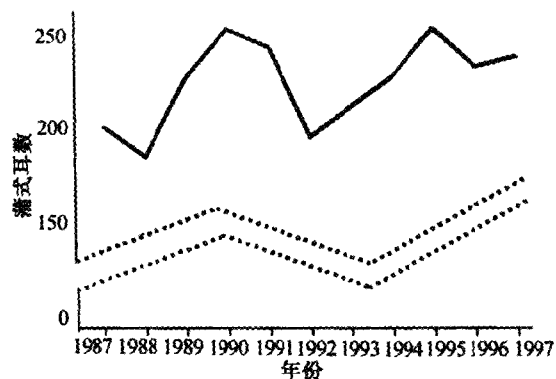


图 1-12

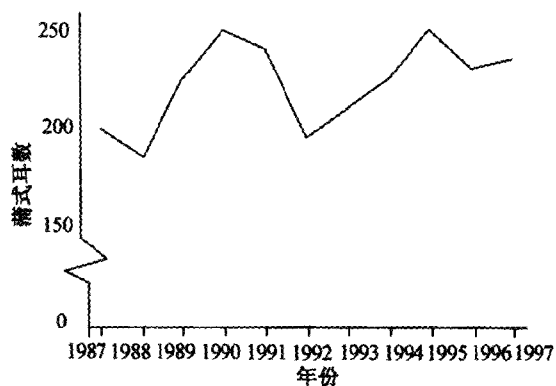


图 1-13

1.27 表1.6给出了美国境内五大湖的面积, 请根据数据作图.

表 1.6

五大湖	面积(平方英里)
密歇根湖	22 342
苏比利尔湖	20 557
休伦湖	8 800
伊利湖	5 033
安大略湖	3 446
总计	60 178

来源: 美国调查局

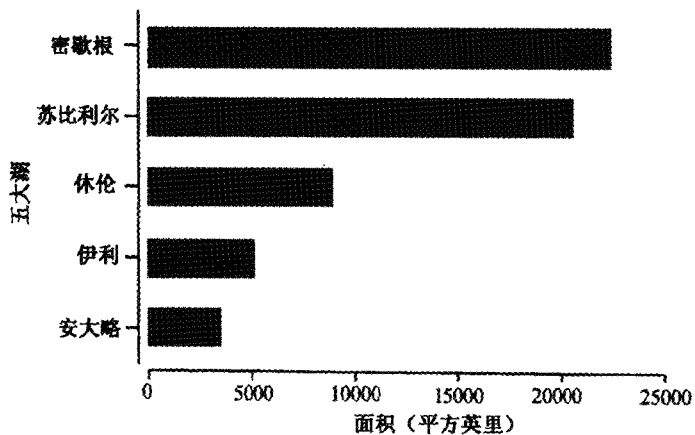


图 1-14 五大湖面积(平方英里)

解 解法一 图1-14是一个水平条形图.按湖的面积由小到大排列.

解法二 图1-15称为**圆形图**.为了作圆形图,我们这样来考虑:总面积60 178平方英里,对应于整个圆的角度数,即 360° .因此,1平方英里对应于 $360^\circ/60\,178$.苏比利尔湖的20 557平方英里对应角度是 $20\,557 \times (360^\circ/60\,178) = 123^\circ$,而密歇根湖、休伦湖、伊利湖和安大略湖对应的角度分别为 134° , 53° , 30° 和 20° .可用量角器作出所需角度.

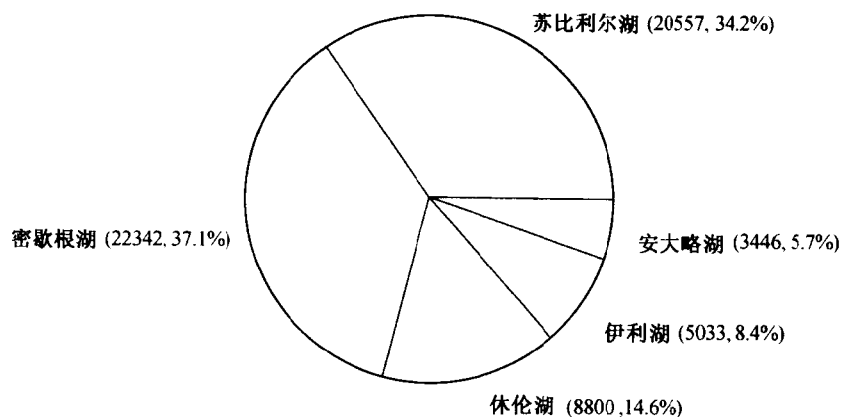


图 1-15 五大湖面积

1.28 长度为 L (厘米)的单摆作一次完整的摆动所需的时间为 T (秒),如表 1.7 所示,这些观察值均出自于物理实验室.

(a) 用图像表示关于 L 的函数 T .

(b) 从(a)的图中,估计单摆长度为 40 cm 时的 T 值.

表 1.7

L	10.1	16.2	22.2	33.8	42.0	53.4	66.7	74.5	86.6	100.0
T	0.64	0.81	0.95	1.17	1.30	1.47	1.65	1.74	1.87	2.01

解 (a)图 1-16 用光滑曲线将观察值点连接了起来.

(b) T 的估计值为 1.27 秒.

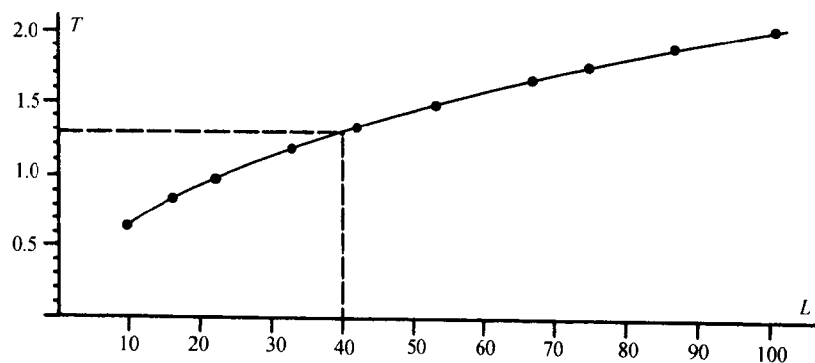


图 1-16

方程

1.29 解下列方程:

(a) $4a - 20 = 8$;

(c) $18 - 5b = 3(b + 8) + 10$;

(b) $3X + 4 = 24 - 2X$;

(d) $\frac{Y+2}{3} + 1 = \frac{Y}{2}$.

解 (a) 两边同时加上 20: $4a - 20 + 20 = 8 + 20$, 即 $4a = 28$.

两边同时除以 4: $4a/4 = 28/4$, 得 $a = 7$.

检验: $4 \times 7 - 20 = 8$, $28 - 20 = 8$, 得 $8 = 8$.

(b) 两边同时减去 4: $3X + 4 - 4 = 24 - 2X - 4$, 即 $3X = 20 - 2X$.

两边同时加上 $2X$: $3X + 2X = 20 - 2X + 2X$, 即 $5X = 20$.

两边同时除以 5: $5X/5 = 20/5$, 得 $X = 4$.

检验: $3 \times 4 + 4 = 24 - 2 \times 4$, $12 + 4 = 24 - 8$, 得 $16 = 16$.

注意, 方程中任何一项可移动, 即从一边移动到另一边, 且符号改变, 称之为移项, 这样方程的求解就方便多了. 因此我们写成:

$$3X + 4 = 24 - 2X \quad 3X + 2X = 24 - 4 \quad 5X = 20 \quad X = 4$$

(c) $18 - 5b = 3b + 24 + 10$, 得 $18 - 5b = 3b + 34$.

移项: $-5b - 3b = 34 - 18$, 即 $-8b = 16$.

除以 -8 : $-8b/(-8) = 16/(-8)$, 得 $b = -2$.

检验: $18 - 5 \times (-2) = 3 \times (-2 + 8) + 10$, $18 + 10 = 3 \times 6 + 10$, 得 $28 = 28$.

(d) 两边同时乘以最小公倍数 6,

$$6\left(\frac{Y+2}{3} + 1\right) = 6\left(\frac{Y}{2}\right) \quad 6\left(\frac{Y+2}{3}\right) + 6 = \frac{6Y}{2} \quad 2(Y+2) + 6 = 3Y$$
$$2Y + 4 + 6 = 3Y \quad 2Y + 10 = 3Y \quad 10 = 3Y - 2Y \quad Y = 10.$$

检验: $\frac{10+2}{3} + 1 = \frac{10}{2}$, $\frac{12}{3} + 1 = \frac{10}{2}$, $4 + 1 = 5$, 得 $5 = 5$.

1.30 解下列联立方程:

$$(a) 3a - 2b = 11$$

$$(b) 5X + 14Y = 78$$

$$(c) 3a + 2b + 5c = 15$$

$$5a + 7b = 39;$$

$$7X + 3Y = -7;$$

$$7a - 3b + 2c = 52$$

$$5a + b - 4c = 2.$$

解 (a) 第一式乘以 7: $21a - 14b = 77$

(1)

第二式乘以 2: $10a + 14b = 78$

(2)

$$\text{相加:} \quad 31a \quad = 155$$

$$\text{除以 31:} \quad a = 5$$

注意, 在给定方程两边同时乘以适当的数, 我们能写出两个等价方程(1)和(2), 在这两式里未知数 b 的系数数值是相等的. 然后通过加法, 我们就能消去未知数 b , 得到 a .

把 $a = 5$ 带入第一个方程: $3 \times 5 - 2b = 11$, $-2b = -4$, 得 $b = 2$. 因此 $a = 5$, $b = 2$.

检验: $3 \times 5 - 2 \times 2 = 11$, $15 - 4 = 11$, $11 = 11$; $5 \times 5 + 7 \times 2 = 39$, $25 + 14 = 39$, $39 = 39$.

$$(b) \text{第一式乘以 3:} \quad 15X + 42Y = 234$$

(3)

$$\text{第二式乘以 -14:} \quad -98X - 42Y = 98$$

(4)

$$\text{相加:} \quad -83X \quad = 332$$

$$\text{除以 -83:} \quad X = -4$$

把 $X = -4$ 代入第一式: $5(-4) + 14Y = 78$, $14Y = 98$, $Y = 7$. 因此 $X = -4$, $Y = 7$.

检验: $5(-4) + 14 \times 7 = 78$, $-20 + 98 = 78$, $78 = 78$; $7(-4) + 3 \times 7 = -7$, $-28 + 21 = -7$, $-7 = -7$.

$$(c) \text{第一式乘以 2:} \quad 6a + 4b + 10c = 30$$

$$\text{第二式乘以 -5:} \quad -35a + 15b - 10c = -260$$

$$\text{相加:} \quad -29a + 19b \quad = -230$$

(5)

$$\text{第二式乘以 2:} \quad 14a - 6b + 4c = 104$$

$$\text{第三式照抄:} \quad 5a + b - 4c = 2$$

$$\text{相加:} \quad 19a - 5b \quad = 106$$

(6)

我们消去了 c , 还剩下两个方程(5)和(6)来同时求出 a 和 b .

$$(5)\text{式乘以 } 5: \quad -145a + 95b = -1150$$

$$(6)\text{式乘以 } 19: \quad 361a - 95b = 2014$$

$$\text{相加:} \quad 216a \quad = 864$$

$$\text{除以 } 216: \quad a = 4$$

把 $a = 4$ 代入(5)或(6), 我们得到 $b = -6$.

把 $a = 4, b = -6$ 代入任何一个给定方程, 我们得到 $c = 3$.

因此 $a = 4, b = -6, c = 3$.

检验: $3 \times 4 + 2(-6) + 5 \times 3 = 15, 15 = 15; 7 \times 4 - 3(-6) + 2 \times 3 = 52, 52 = 52; 5 \times 4 + (-6) - 4 \times 3 = 2, 2 = 2$.

不等式

1.31 用语言表达下列各式含义:

$$(a) N > 30; \quad (b) X \leq 12; \quad (c) 0 < p \leq 1; \quad (d) \mu - 2t < X < \mu + 2t.$$

解 (a) N 比 30 大.

(b) X 小于或等于 12.

(c) p 比 0 大但小于或等于 1.

(d) X 比 $\mu - 2t$ 大但小于 $\mu + 2t$.

1.32 用符号表示下列各式:

(a) X 的值可介于 2 和 5 之间, 并可等于 2 和 5;

(b) 算术平均值 \bar{X} 比 28.42 大, 但比 31.56 小;

(c) 正数 m 小于等于 10;

(d) P 是一非负数.

解 (a) $2 \leq X \leq 5$; (b) $28.42 < \bar{X} < 31.56$; (c) $0 < m \leq 10$; (d) $P \geq 0$.

1.33 按要求用不等号排列数字 3.42, -0.6 , -2.1 , 1.45 和 -3 .

(a) 按大小升序排列;

(b) 按大小降序排列.

解 (a) $-3 < -2.1 < -0.6 < 1.45 < 3.42$;

(b) $3.42 > 1.45 > -0.6 > -2.1 > -3$.

这些数字作为点画在一直线上(见习题 1.18), 它们从左至右逐渐增大.

1.34 在下列各式中, 找出 X 的对应不等式(即解 X 的不等式):

$$(a) 2X < 6; \quad (c) 6 - 4X < -2; \quad (e) -1 \leq \frac{3-2X}{5} \leq 7.$$

$$(b) 3X - 8 \geq 4; \quad (d) -3 < \frac{X-5}{2} < 3;$$

解 (a) 两边同时除以 2, 得 $X < 3$.

(b) 两边同时加 8, $3X \geq 12$; 两边同时除以 3, $X \geq 4$.

(c) 两边同时加 -6 , $-4X < -8$; 两边同时除以 -4 , $X > 2$. 注意, 在不等式里, 我们可以像在等式中一样, 通过改变符号把一项从不等式的一边移到另一边. 比如, 在(b)中, $3X \geq 4 + 8$.

(d) 两边同时乘以 2, $-6 < X - 5 < 6$; 两边同时加 5, $-1 < X < 11$.

(e) 两边同时乘以 5, $-5 \leq 3 - 2X \leq 35$; 两边同时加 -3 , $-8 \leq -2X \leq 32$; 两边同时除以 -2 , $4 \geq X \geq -16$, 即 $-16 \leq X \leq 4$.

对数和反对数

1.35 指出下列数的常用对数(以 10 为底)的首数:

- (a) 57; (d) 35.63; (g) 186 000; (j) 0.0325;
(b) 57.4; (e) 982.5; (h) 0.71; (k) 0.0071;
(c) 5.63; (f) 7824; (i) 0.7314; (l) 0.0003.

解 (a) 1; (b) 1; (c) 0; (d) 1; (e) 2; (f) 3; (g) 5; (h) $9 - 10$; (i) $9 - 10$; (j) $8 - 10$; (k) $7 - 10$; (l) $6 - 10$.

1.36 求对数:

- (a) $\log 87.2$; (f) $\log 0.382$; (k) $\log 4.638$; (p) $\log 0.2548$
(b) $\log 37\ 300$; (g) $\log 0.00159$; (l) $\log 6.753$; (q) $\log 0.04372$;
(c) $\log 753$; (h) $\log 0.0753$; (m) $\log 183.2$; (r) $\log 0.009848$;
(d) $\log 9.21$; (i) $\log 0.000827$; (n) $\log 43.15$; (s) $\log 0.0001788$.
(e) $\log 54.50$; (j) $\log 0.0503$; (o) $\log 876\ 400$;

解 (a) 尾数 = .9405, 首数 = 1, 因此 $\log 87.2 = 1.9405$;

(b) 4.5717

(c) 2.8768

(d) 0.9643

(e) 1.7364

(f) 尾数 = .5821, 首数 = $9 - 10$; 因此 $\log 0.382 = 9.5821 - 10$.

(g) $7.2014 - 10$

(h) $8.8768 - 10$

(i) $6.9175 - 10$

(j) $8.7016 - 10$

(k) $\log 4638$ 的尾数是 $\log 4630$ 和 $\log 4640$ 尾数差的 0.8 倍.

$\log 4640$ 的尾数 = .6665

$\log 4630$ 的尾数 = .6656

差 = .0009

$\log 4.638$ 的尾数 = $.6656 + 0.8 \times 0.0009 = .6663$; 因此 $\log 4.638 = .6663$. 这个过程称为线性插值. 如需要, 附录 VII 的相应内容可直接用来求出尾数 ($.6656 + 7$).

(l) $0.8295(8293 + 2)$

(m) $2.2630(2625 + 5)$

(n) $1.6350(6345 + 5)$

(o) $5.9427(9425 + 2)$

(p) $9.4062 - 10(4048 + 14)$

(q) $8.6407 - 10(6405 + 2)$

(r) $7.9933 - 10(9930 + 3)$

(s) $6.2524 - 10(2504 + 20)$

1.37 求反对数:

- (a) $\text{antilog } 1.9058$; (c) $\text{antilog } 7.8657 - 10$; (f) $\text{antilog } 2.6715$
(b) $\text{antilog } 3.8531$ (d) $\text{antilog } 9.8267 - 10$ $\text{antilog } 4.1853$
 $\text{antilog } 2.1875$ $\text{antilog } 2.3927$ $\text{antilog } 0.9245$;
 $\text{antilog } 0.4997$ $\text{antilog } 7.7443 - 10$; (g) $\text{antilog } 1.6089$
 $\text{antilog } 4.9360$; (e) $\text{antilog } 9.3842 - 10$; $\text{antilog } 8.8907 - 10$
 $\text{antilog } 1.2000$.

解 (a) 附录 VII 中, .9058 的尾数对应数字是 805. 由于首数是 1, 所求数小数点前一定有两位数,

因此所求的数为 80.5(即, $\text{antilog}1.9058 = 80.5$).

(b) $\text{antilog}3.8531 = 7130$, $\text{antilog}2.1875 = 154$, $\text{antilog}0.4997 = 3.16$, $\text{antilog}4.9360 = 86\,300$.

(c)附录Ⅶ中, .8657 的尾数对应数字是 734. 由于首数是 7-10, 所求数小数点后直接跟有两个零. 因此所求的数为 0.00734(即, $\text{antilog}7.8657 - 10 = 0.00734$). 可查阅附录Ⅶ中的相关内容.

(d) $\text{antilog}9.8267 - 10 = 0.671$, $\text{antilog}2.3927 = 0.0247$, $\text{antilog}7.7443 - 10 = 0.00555$.

(e)由于在表中找不到相应尾数, 必须使用插值法:

$\log 2430$ 的尾数 = .3856 给定尾数 = .3824

$\log 2420$ 的尾数 = .3838 下一个更小的尾数 = .3838

差 = .0018

差 = .0004

因此 $2420 + (4/18)(2430 - 2420) = 2422$, 所求数为 0.2422.

(f) $\text{antilog}2.6715 = 469.3(3/9 \times 10 \approx 3)$, $\text{antilog}4.1853 = 15\,320(6/28 \times 10 \approx 2)$, $\text{antilog}0.9245 = 8.404(2/5 \times 10 = 4)$.

(g) $\text{antilog}1.6089 = 0.4064(4/11 \times 10 \approx 4)$, $\text{antilog}8.8907 - 10 = 0.07775(3/6 \times 10 = 5)$, $\text{antilog}1.2000 = 15.85(13/27 \times 10 \approx 5)$.

用对数计算

用对数计算下列各式的值.

1.38 $P = 3.81 \times 43.4$.

解 $\log P = \log 3.81 + \log 43.4$ $\log 3.81 = 0.5809$

(+) $\log 43.4 = 1.6375$

$\log P = 2.2184$

因此 $P = \text{antilog}2.2184 = 165.3$, 或保留三个有效数字得 165. 注意指数形式的运算:

$$3.81 \times 43.4 = 10^{0.5809} \times 10^{1.6375} = 10^{0.5809+1.6375} = 10^{2.2184} = 165.3$$

1.39 $P = 73.42 \times 0.004620 \times 0.5143$.

解 $\log P = \log 73.42 + \log 0.004620 + \log 0.5143$

$\log 73.42 = 1.8658$

(+) $\log 0.004620 = 7.6646 - 10$

(+) $\log 0.5143 = 9.7112 - 10$

$\log P = 19.2416 - 20 = 9.2416 - 10$

因此 $P = 0.1744$.

1.40 $P = \frac{784.6 \times 0.0431}{28.23}$.

解 $\log P = \log 784.6 + \log 0.0431 - \log 28.23$

$\log 784.6 = 2.8947$

(+) $\log 0.0431 = 8.6345 - 10$

$11.5292 - 10$

(-) $\log 28.23 = 1.4507$

$\log P = 10.0785 - 10 = 0.0785$

因此 $P = 1.198$, 或保留三个有效数字得 1.20. 注意指数形式的运算:

$$\begin{aligned} \frac{784.6 \times 0.0431}{28.23} &= \frac{10^{2.8947} \times 10^{8.6345-10}}{10^{1.4507}} = 10^{2.8947+8.6345-10-1.4507} \\ &= 10^{0.0785} = 1.198 \end{aligned}$$

$$1.41 \quad P = (5.395)^8.$$

$$\text{解} \quad \log P = 8 \log 5.395 = 8 \times 0.7320 = 5.8560, P = 717800 \text{ 或 } 7.178 \times 10^5.$$

$$1.42 \quad P = \sqrt{387.2} = (387.2)^{1/2}.$$

$$\text{解} \quad \log P = \frac{1}{2} \log 387.2 = \frac{1}{2} \times 2.5879 = 1.2940, P = 19.68.$$

$$1.43 \quad P = (0.08317)^{1/5}.$$

$$\text{解} \quad \log P = \frac{1}{5} \log 0.08317 = \frac{1}{5} \times (8.9200 - 10) = \frac{1}{5} \times (48.9200 - 50) = 9.7840 - 10, \\ P = 0.6081.$$

$$1.44 \quad P = \frac{\sqrt{0.003654} \times 18.37^3}{8.724^4 \times \sqrt[4]{743.8}}.$$

$$\text{解} \quad \log P = \frac{1}{2} \log 0.003654 + 3 \log 18.37 - (4 \log 8.724 + \frac{1}{4} \log 743.8)$$

分子 N

分母 D

$$\frac{1}{2} \log 0.003654 = \frac{1}{2} (7.5628 - 10)$$

$$4 \log 8.724 = 4 \times 0.9407 = 3.7628$$

$$= \frac{1}{2} (17.5628 - 20) = 8.7814 - 10$$

$$\frac{1}{4} \log 743.8 = \frac{1}{4} \times 2.8714 = 0.7178$$

$$3 \log 18.37 = 3 \times 1.2641 = 3.7923$$

$$\text{相加:} \quad \log N = 12.5737 - 10$$

$$\text{相加:} \quad \log D = 4.4806$$

$$(-) \log D = 4.4806$$

$$\log P = 8.0931 - 10$$

$$P = 0.01239$$

$$1.45 \quad P = \sqrt{\frac{874.3 \times 0.03816 \times 28.53^3}{1.754^4 \times 0.007352}}.$$

$$\text{解} \quad \log P = \frac{1}{2} [\log 874.3 + \log 0.03816 + 3 \log 28.53 - (4 \log 1.754 + \log 0.007352)]:$$

$$\log 874.3 = 2.9417 = 2.9417$$

$$\log 0.03816 = 8.5816 - 10 = 8.5816 - 10$$

$$3 \log 28.53 = 3 \times 1.4553 = 4.3659$$

$$\text{相加:} \quad 15.8892 - 10 \quad (1)$$

$$4 \log 1.754 = 4 \times 0.2440 = 0.9760$$

$$\log 0.007352 = 7.8664 - 10$$

$$\text{相加:} \quad 8.8424 - 10 \quad (2)$$

从(1)和(2)我们得到:

$$\log P = \frac{1}{2} [(15.8892 - 10) - (8.8424 - 10)] = \frac{1}{2} \times 7.0468 = 3.5234, P = 3338.$$

补充习题

变量

1.46 指出下列哪些是离散数据,哪些是连续数据:

- (a) 一个城市一年中不同月份的降雨量;
- (b) 一辆汽车的速度(英里/小时);
- (c) 任意时刻 20 美元面值的货币在美国流通的数量;
- (d) 股票市场每天抛售的股票总值;
- (e) 几年内一所大学的学生入学人数.

1.47 给出下列变量的值域,并指出哪些是连续变量,哪些是离散变量.

- (a) 一农场几年内每英亩小麦产量 W ;
- (b) 一个家庭的成员人数 N ;
- (c) 一个人的婚姻状况;
- (d) 一枚导弹飞行的速度 T ;
- (e) 一朵花的花瓣数 P .

数据舍入, 科学记数和有效数字

1.48 按要求对下列数据进行舍入:

- (a) 3256 舍入至最近的百位; (f) 3 502 378 舍入至最近的百万位;
- (b) 5.781 舍入至最近的十分位; (g) 148.475 舍入至最近的整数位;
- (c) 0.0045 舍入至最近的千分位; (h) 0.000098501 舍入至最近的百万分位;
- (d) 46.7385 舍入至最近的百分位; (i) 2184.73 舍入至最近的十位;
- (e) 125.9995 保留两位小数; (j) 43.87500 舍入至最近的百分位;

1.49 不用 10 的幂形式表示下列数字:

- (a) 132.5×10^4 ; (c) 280×10^{-7} ; (e) 3.487×10^{-4} ;
- (b) 418.72×10^{-5} ; (d) 7300×10^6 ; (f) 0.0001850×10^5 .

1.50 假定下列数字是精确记录的, 在每个数里有几个有效数字?

- (a) 2.54 厘米; (d) 3.51×10^6 蒲式耳; (g) 378 盎司; (j) 100.00 英里;
- (b) 0.004500 码; (e) 10.000100 英尺; (h) 4.50×10^{-3} 千米;
- (c) 3 510 000 蒲式耳; (f) 378 人; (i) 500.8×10^5 千克.

1.51 假定下列数字是精确记录的, 在每个数里有几个有效数字? 每个数的最大误差是多少?

- (a) 7.20×10^6 蒲式耳; (c) 5280 英尺; (e) 186 000 英里/秒;
- (b) 0.00004835 厘米; (d) 3.0×10^8 米; (f) 186 千英里/秒.

1.52 用科学记数法书写下列数字(除非另外说明, 否则认为数字是精确记录的).

- (a) 0.000317; (d) 0.000009810;
- (b) 428 000 000(四个有效数字); (e) 732 个千;
- (c) 21 600.00; (f) 18.0 的千分之十.

数值计算

1.53 假设 72.48 和 5.16 分别有 4 个和 3 个有效数字, 说明它们的(a)乘积;(b)商不能精确到超过 3 个有效数字. 写出精确的乘积和商.

1.54 完成下列运算.

- (a) 0.36×781.4 ; (b) $\frac{873.00}{4.881}$; (c) $5.78 \times 2700 \times 16.00$; (d) $\frac{0.00480 \times 2300}{0.2084}$;
- (e) $\sqrt{120 \times 0.5386 \times 0.4614}$ (120 是准确的); (f) $\frac{416\,000 \times 0.000187}{\sqrt{73.84}}$;
- (g) $14.8641 + 4.48 - 8.168 + 0.36125$; (h) $4\,173\,000 - 170\,264 + 1\,820\,470 - 78\,320$ (数据分别精确到 4 个, 6 个, 6 个, 5 个有效数字);
- (i) $\sqrt{\frac{7 \times 4.386^2 - 3 \times 6.47^2}{6}}$ (3, 6 和 7 是准确的);
- (j) $4.120 \sqrt{\frac{3.1416 \times (9.483^2 - 5.075^2)}{0.0001980}}$.

1.55 当 $U = -2$, $V = 1/2$, $W = 3$, $X = -4$, $Y = 9$, $Z = 1/6$ 时, 求下列各式的值. 假设所有数据是准确的.

- (a) $4U + 6V - 2W$; (f) $3X(4Y + 3Z) - 2Y(6X - 5Z) - 25$;
- (b) $\frac{XYZ}{UVW}$; (g) $\sqrt{\frac{(W-2)^2}{V} + \frac{(Y-5)^2}{Z}}$;
- (c) $\frac{2X-3Y}{UW+XV}$; (h) $\frac{X-3}{\sqrt{(Y-4)^2 + (U+5)^2}}$;
- (d) $3(U-X)^2 + Y$; (i) $X^3 + 5X^2 - 6X - 8$;
- (e) $\sqrt{U^2 - 2UV + W}$; (j) $\frac{U-V}{\sqrt{U^2 + V^2}} [U^2 V (W+X)]$.

函数, 表和图

- 1.56 变量 X 和 Y 的关系由方程 $Y = 10 - 4X$ 决定.
- (a) 当 $X = -3, -2, -1, 0, 1, 2, 3, 4$ 和 5 时, 求 Y . 并把结果列在表中;
- (b) 当 $X = -2.4, -1.6, -0.8, 1.8, 2.7, 3.5$ 和 4.6 时, 求 Y ;
- (c) 如果 X 和 Y 的关系由方程 $Y = F(X)$ 表示, 求 $F(2.8), F(-5), F(\sqrt{2})$ 和 $F(-\pi)$;
- (d) 当 $Y = -2, 6, -10, 1.6, 16, 0$ 和 10 时, 求相应的 X ;
- (e) 用 Y 来表示 X .
- 1.57 如果 $Z = X^2 - Y^2$, 当 (a) $X = -2, Y = 3$; (b) $X = 1, Y = 5$ 时, 求 Z ; (c) 如用函数记法 $Z = F(X, Y)$, 求 $F(-3, -1)$.
- 1.58 如果 $W = 3XZ - 4Y^2 + 2XY$, 当 (a) $X = 1, Y = -2, Z = 4$; (b) $X = -5, Y = -2, Z = 0$ 时, 求 W ; (c) 如用函数记法 $W = F(X, Y, Z)$, 求 $F(3, 1, -2)$.
- 1.59 在直角坐标中标出下列各点: (a) $(3, 2)$, (b) $(2, 3)$, (c) $(-4, 4)$, (d) $(4, -4)$, (e) $(-3, -2)$, (f) $(-2, -3)$, (g) $(-4.5, 3)$, (h) $(-1.2, -2.4)$, (i) $(0, -3)$, (j) $(1.8, 0)$.
- 1.60 作出下列方程的图像 (a) $Y = 10 - 4X$ (参见习题 1.56); (b) $Y = 2X + 5$; (c) $Y = \frac{1}{3}(X - 6)$, (d) $2X + 3Y = 12$, (e) $3X - 2Y = 6$.
- 1.61 作出下列方程的图像 (a) $Y = 2X^2 + X - 10$, (b) $Y = 6 - 3X - X^2$.
- 1.62 作 $Y = X^3 - 4X^2 + 12X - 6$ 的图像.
- 1.63 表 1.8 给出了从 1989 到 1995 年间, 男性和女性艾滋病患者死亡人数. 在同一坐标系中根据数据作出两条相应的线图.

表 1.8

年份	1989	1990	1991	1992	1993	1994	1995
男性	23 742	26 752	30 725	34 072	35 551	37 360	26 375
女性	2 613	3 182	3 926	4 741	5 526	6 615	4 881

来源: 美国疾病控制中心.

- 1.64 利用表 1.8 中的数据, 建立与图 1-9 和图 1-10 相似的条形图.
- 1.65 根据习题 1.63 的表 1.8, 写出男性和女性艾滋病患者年死亡人数占所有艾滋病死亡人数的百分比. 根据这些百分比作出百分数分支图.
- 1.66 表 1.9 给出了美国 1990 到 1994 年间白人和有色人种每 1000 个新生儿的婴儿死亡率. 用适当的图形来描绘数据情况.

表 1.9

年份	1990	1991	1992	1993	1994
白人	7.6	7.3	6.9	6.8	6.6
有色人种	15.5	15.1	14.4	14.1	13.5

来源: 美国国家健康统计中心, 美国生死统计.

- 1.67 表 1.10 给出了太阳系中行星的轨道行驶速度(英里/秒), 根据数据作图.

表 1.10

行星	水星	金星	地球	火星	木星	土星	天王星	海王星	冥王星
速度	29.7	21.8	18.5	15.0	8.1	6.0	4.2	3.4	3.0

1.68 表 1.11 给出了 2000~2006 年公立学校从 K 到 8 年级,9~12 年级及大学计划入学人数(以千记).用线图、条形图和分支条形图作图.

表 1.11

年份	2000	2001	2002	2003	2004	2005	2006
K~8 年级	33 852	34 029	34 098	34 065	33 882	33 680	33 507
9~12 年级	13 804	13 862	14 004	14 169	14 483	14 818	15 021
大学	12 091	12 225	12 319	12 420	12 531	12 646	12 768

来源:美国国家教育及规划统计中心,年鉴.

1.69 根据表 1.11 的数据作百分数分支图.

1.70 表 1.12 显示美国 1995 年男性和女性(18 岁以上)的婚姻状况.根据数据,(a)用同一直径作两个相应圆形图,(b)作一你任选的图形.

表 1.12

婚姻状况	男性(占总数百分比)	女性(占总数百分比)
未婚	26.8	19.4
已婚	62.7	59.2
丧偶	2.5	11.1
离婚	8.0	10.3

来源:美国人口普查局——最新人口报告.

1.71 表 1.13 给出了 1987~1994 年美国破产申请文件总数.根据所给数据,作出相应类型的图.

表 1.13

年份	1987	1988	1989	1990	1991	1992	1993	1994
破产申请文件总数	561 278	594 567	642 993	725 484	880 399	972 490	918 734	845 257

来源:美国法院行政管理办公室,主任年度汇报.

1.72 表 1.14 显示美国 1988~1995 年,每 100 000 个居民中的犯罪率.根据数据,作两种类型的图.

表 1.14

年份	1988	1989	1990	1991	1992	1993	1994	1995
每 100 000 个居民中的犯罪率	5 664.2	5 741.0	5 820.3	5 897.8	5 660.2	5 484.4	5 373.5	5 277.6

来源:美国联邦调查局——美国人的犯罪情况——1995.

1.73 表 1.15 显示了 1997 年 7 个人口最多的国家的人口数.用圆形图来说明 7 个人口最多国家的人口情况.

表 1.15

国家	中国	印度	美国	印度尼西亚	巴西	俄罗斯	巴基斯坦
人口数(百万)	1 222	968	268	210	165	148	132

来源:美国人口普查局,国际资料总部.

1.74 Pareto 图中的条带是根据频率值而决定的. 因此最高的条带在左边, 最低的条带在右边. 为表 1.15 的数据建立一个 Pareto 图.

1.75 表 1.16 显示了世界上各大洋的面积(百万平方英里). 根据数据, 做(a)条形图; (b)圆形图.

表 1.16

大洋	太平洋	大西洋	印度洋	南极洲	北冰洋
面积	63.8	31.5	28.4	7.6	4.8

来源: 联合国.

方程

1.76 解下列方程:

- (a) $16 - 5c = 36$; (d) $3(2U + 1) = 5(3 - U) + 3(U - 2)$;
 (b) $2Y - 6 = 4 - 3Y$; (e) $3[2(X + 1) - 4] = 10 - 5(4 - 2X)$;
 (c) $4(X - 3) - 11 = 15 - 2(X + 4)$; (f) $(2/5)(12 + Y) = 6 - (1/4)(9 - Y)$.

1.77 解下列联立方程:

- (a) $2a + b = 10$ (e) $2a + b - c = 2$
 $7a - 3b = 9$; $3a - 4b + 2c = 4$
 $4a - 3b - 5c = -8$;
 (b) $3a + 5b = 24$ (f) $5X + 2Y + 3Z = -5$
 $2a + 3b = 14$; $2X - 3Y - 6Z = 1$
 (c) $8X - 3Y = 2$ $X + 5Y - 4Z = 22$;
 $3X + 7Y = -9$; (g) $3U - 5V + 6W = 7$
 (d) $5A - 9B = -10$ $5U + 3V - 2W = -1$
 $3A - 4B = 16$; $4U - 8V + 10W = 11$.

1.78 (a) 在同一坐标系中做出方程 $5X + 2Y = 4$ 和 $7X - 3Y = 23$ 的图像;

(b) 从图像中求联立方程的解;

(c) 用(a)和(b)的方法求习题 1.77 中(a)~(d)的解.

1.79 (a) 用习题 1.61(a)的图像解方程 $2X^2 + X - 10 = 0$ (提示: 抛物线与 X 轴的交点的横坐标即为方程的解);

(b) 用(a)的方法解 $3X^2 - 4X - 5 = 0$.

1.80 二次方程 $aX^2 + bX + c = 0$ 解的公式为 $X = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$. 用公式解(a) $3X^2 - 4X - 5 = 0$, (b) $2X^2 + X - 10 = 0$, (c) $5X^2 + 10X = 7$, (d) $X^2 + 8X + 25 = 0$.

不等式

1.81 按要求排列数字 -4.3 , -6.15 , 2.37 , 1.52 和 -1.5 . (a) 升序排列; (b) 降序排列.

1.82 用不等号表达下列各式:

- (a) 学生数 N 介于 $30 \sim 50$ 之间(包括 30 及 50);
 (b) 一对骰子的点数之和 S 不小于 7 ;
 (c) X 大于等于 -4 但小于 3 ;
 (d) P 最多为 5 ;
 (e) X 至少比 Y 大 2 .

1.83 解下列不等式:

- (a) $3X \geq 12$; (e) $-3 \leq \frac{1}{5}(2X + 1) \leq 3$;
 (b) $4X < 5X - 3$; (f) $0 < \frac{1}{2}(15 - 5N) \leq 12$;
 (c) $2N + 15 > 10 + 3N$; (g) $-2 \leq 3 + \frac{1}{2}(a - 12) < 8$.
 (d) $3 + 5(Y - 2) \leq 7 - 3(4 - Y)$;

对数和反对数

1.84 求下列各数的常用对数:

- (a)387; (c)0.0792; (e)0.6042; (g)476.3; (i)7.146; (k)0.00098;
(b)0.387; (d)14 630; (f)0.002795; (h)1.007; (j)71.46; (l)84 620 000.

1.85 求下列各数的反对数:

- (a)3.5611; (c)1.7045; (e)2.4700; (g) $\bar{2}.8003$; (i)0.0800;
(b)9.8293 - 10; (d)8.9266 - 10; (f)6.4700 - 10; (h)3.7072; (j)6.3841.

1.86 用对数计算下列各式的值:

- (a)783.6 × 1654; (f)0.04182 × $\sqrt{0.6758}$;
(b) $\frac{21.7}{378.2}$; (g) $\sqrt[3]{3728}$;
(c) $\frac{0.04556 \times 624.1}{14.32 \times 0.003572}$; (h) $\sqrt[5]{21.63 \times 33.81 \times 47.53 \times 65.28 \times 87.47}$;
(d)1.562¹⁵; (i) $\sqrt{\frac{48.79 \times 0.00574^3}{2.143^5}}$;
(e) $\frac{0.3854^4 \times 12.48^2}{0.04382^3}$; (j) $\frac{3.781}{0.01873} \sqrt{\frac{43.25 \times 0.08743}{0.002356 \times 6.824}}$.

1.87 作出(a) $Y = \log X$ 和(b) $Y = 10^X$ 的图像,并讨论它们的相似性.

1.88 改写下列方程,使之不出现对数:(a) $2\log X - 3\log Y = 2$, (b) $\log Y + 2X = \log 3$.

1.89 如 $a^p = N$, 其中 a 和 p 是正数且 $a \neq 1$, 我们称 p 是以 a 为底 N 的对数, 记作 $p = \log_a N$. 求 (a) $\log_2 8$,
(b) $\log_{25} 125$, (c) $\log_4 1/16$, (d) $\log_{1/2} 32$, (e) $\log_5 1$.

1.90 证明:近似地, $\log_e N = 2.303 \log_{10} N$, 其中 $e = 2.71828\cdots$ 称为对数的自然基底, $N > 0$.

1.91 证明 $(\log_a a)(\log_a b) = 1$, 其中 $a > 0$, $b > 0$, $a \neq 1$, $b \neq 1$.

第二章 频数分布

原始数据

原始数据是收集来的没经过整理的数据. 比如从一个大学按字母排列的名单上抽取到的 100 个男同学的身高是一组原始数据.

数组阵列

数组阵列是原始数据按数量大小升序或降序排列的序列. 数据中最大值与最小值的差称为数据的**全距**. 例如, 如果 100 个男同学中身高最高为 74 英寸, 最低为 60 英寸, 则全距为 $74 - 60 = 14$ 英寸.

频数分布

当汇总大量的原始数据时, 把数据按**类型分组**会带来方便. 其中每个组的数据个数, 称为**组频数**. 表示各组及它们对应的组频数的表格称为**频数分布**或**频数表**. 表 2.1 就显示了 XYZ 大学 100 个男同学身高的频数分布(记录至最近的英寸).

表 2.1 XYZ 大学 100 个男同学的身高

身高(英寸)	学生数
60~62	5
63~65	18
66~68	42
69~71	27
72~74	8
	总数 100

例如, 第一组由 60~62 英寸的身高构成, 用符号 60~62 表示. 由于有 5 个学生的身高属于这一组, 因此这一组的组频数是 5.

在上述频数分布中, 经过整理和汇总的数据称为**分类资料**. 尽管在归组过程中数据的许多最初细节已被改变, 然而我们却对数据的整体情况有了清楚的了解, 而且数据之间的相互关系也一目了然.

组距和组限

用来定义某一组的符号, 称为**组距**, 如表 2.1 中 60~62. 两端的数 60 和 62, 称为**组限**, 其中较小的数(60)称为**下组限**, 较大的数(62)则称为**上组限**. 尽管组距实际上只是组的符号, 但**组**和**组距**在使用中常可互相代替.

从理论上说, 一个没有上组限或下组限的组距称为**开组距**. 例如, 考虑成年人的年龄, 组距“65 及 65 岁以上”就是一个开组距.

组界

如果身高记录至最近的英寸, 那么组距 60~62 理论上包括了从 59.5000 到 62.5000 英寸的所有测量值. 这些数简单地记为精确数 59.5 和 62.5, 并称为**组界**或**真实组限**, 其中较小的

数(59.5)称为**下组界**,较大的数(62.5)称为**上组界**.

在实际中,组界可以由一个组距中的上组限和较高一级组距中的下组限相加除以 2 而得.

有时也用组界来标记不同的组.比如,表 2.1 中第一列的不同组就可用 59.5~62.5, 62.5~65.5 等等来表示.为了避免在使用中出现意义不明确的情况,组界不应与实际观察值一致.因为如果一个观察值为 62.5,那么就不容易判断是属于组距 59.5~62.5,还是属于组距 62.5~65.5.

组距的大小或宽度

组距的大小或宽度是上下组界的差,也常称为**组宽**.如果一个频数分布的所有组距都有同样的宽度,那么这个共同的宽度用 c 来表示.在这种情况下, c 等于 2 个连续下组界或 2 个连续上组界的差.在表 2.1 中, $c = 62.5 - 59.5 = 65.5 - 62.5 = 3$.

组中值

组中值是组距的中点,可以由上下组限的和除以 2 得到.因此,60~62 的组中值为 $(60 + 62)/2 = 61$.组中值也称为**组中点**.

为了深入地进行数学研究,我们常假定一个给定组距的所有观察值都与组中值是一致的.因此,组距 60~62 英寸中所有身高都视为 61 英寸.

建立频数分布的一般法则

1. 找出原始数据中的最大值和最小值,并且求出全距(即最大值与最小值的差).
2. 把全距按组的宽度一致原则恰当地分组.如果这样不可行,那么就分成宽度不同的组或开组距(见习题 2.12).组距数目根据数据情况通常取 5~20 之间.选择组距时,也要注意组中值(或组中点)与实际观察数据应一致.这是保证在进一步的数学研究中,所谓的**分组误差**减少到最小.然而,组界不应与实际观察数据一致.
3. 求出落入每个组距中的观察值数目,即求出组频数.这些最好用**计数**完成(见习题 2.8).

直方图和频数多边形

直方图和频数多边形是频数分布的两种图表表示形式.

1. 一个**直方图**或**频数直方图**由一组满足以下条件的矩形构成:(a)以水平轴(X 轴)为底,中心在组中值且宽度等于组距宽度;(b)面积大小与组频数成比例.

如果所有组距都有同样的宽度,那么矩形的高与组频数成比例,习惯上,就把高度视为组频数.如果组距的宽度不同,那么就要调整矩形的高度(见习题 2.13).

2. **频数多边形**是关于组频数的线形图,依据组中值而得.在直方图中把相邻两矩形上底中点用直线连接起来就可以得到频数多边形.

根据表 2.1 中身高的频数分布而相应作出的直方图或频数多边形,如图 2-1 所示.习惯上,常加上延长部分 PQ 和 RS,其中 Q 和 S 分别为第一组之前一组的组中值和最后一组之后一组的组中值,而它们相应的组频数为 0.在这种情况下,直方图中矩形的面积和等于频数多边形和 X 轴围成区域的总面积(见习题 2.11).

频率分布

一个组的**频率**(相对频数)或**百分率频数**是这个组的频数除以所有组的总频数而得的数值,通常用百分数表示.例如,表 2.1 中 65~68 这一组的频率为 $42/100 = 42\%$.显然,所有组的频率之和为 1,或 100%.

如果表 2.1 中的频数用相应的频率来替换,则所得的表称为**频率分布**,**百分率分布**或**频率**

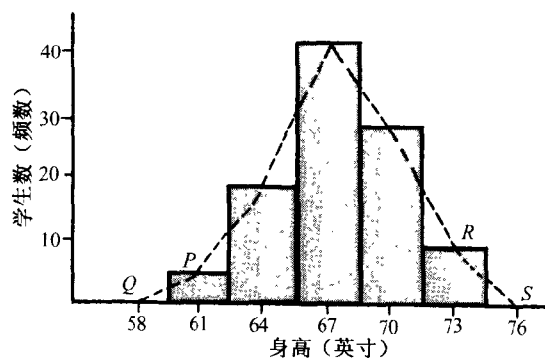


图 2-1

表.

只要在竖轴上把频数改为频率且保持图不变,就可以从直方图或频数多边形中得到频率分布的图像表示.所得的图像分别称为**频率直方图**(或**百分率直方图**)和**频率多边形**(或**百分率多边形**).

累积频数分布和卵形线

一个给定组距中所有小于其上组界的值的总频数称为直到且包括此组距的**累积频数**.例如,表 2.1 中直到且包括组距 66~68 英寸的累积频数为 $5 + 18 + 42 = 65$,这意味着有 65 个学生的身高低于 68.5 英寸.

描绘累积频数的表称为**累积频数分布**或**累积频数表**,简称为**累积分布**,表 2.2 即是表 2.1 中学生身高的累积频数表.

依据上组界来描绘的累积频数低于上组界的图形称为**累积频数多边形**或**卵形线**,如图2-2所示.

表 2.2	
身高(英寸)	学生数
小于 59.5	0
小于 62.5	5
小于 65.5	23
小于 68.5	65
小于 71.5	92
小于 74.5	100

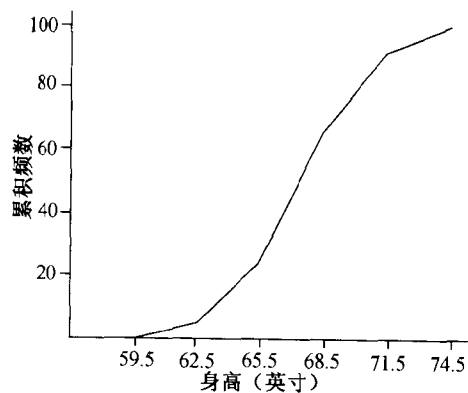


图 2-2

有时为了某种应用,需要考虑所有大于或等于各组下组界的值的累积频数分布.因为在这些应用中,我们考虑 59.5 英寸或更高,62.5 英寸或更高等等,有时这被称为“**不低于**”累积分布,而前面考虑的是“**低于**”累积分布.从一个出发很容易就得到另一个(见习题 2.15).对应的卵形线称为“不低于”和“低于”卵形线.若没有特别声明,一般指的是“低于”类型.

累积频率分布和百分率卵形线

累积频率或**百分率累积频数**,由累积频数除以总频数而得.例如,身高低于 68.5 英寸的累

积频率为 $65/100 = 65\%$,也就是说,65%的学生身高低于68.5英寸.

在表 2.2 和图 2-2 中用累积频率来替代累积频数,就能分别得到**累积频率分布**(或**百分率累积分布**)和**累积频率多边形**(或**百分率卵形线**).

频数曲线和光滑卵形线

收集到的数据通常认为是从较大总体中抽取一个样本获得的.由于总体里有许多可观察的值,从理论上说(对于连续数据)可以选择较小的组距并且仍然有相当多的观察值落入每个组距中.因此对于较大总体我们希望将频数多边形或频率多边形的折线段近似地连成曲线,我们分别称之为**频数曲线**或**频率曲线**.

理论上的这种曲线可通过使样本的频数多边形或频率多边形变光滑而近似得到,样本容量增加,近似程度也将得到改善.也由于这个原因,频数曲线有时也称为**光滑频数多边形**.

同样地,使累积频数多边形或卵形线变光滑就可以得到**光滑卵形线**.通常,使卵形线变光滑要比使频数多边形变光滑要容易些(见习题 2.18).

频数曲线的种类

在实际中产生的频数曲线呈现出某些特有的形状,如图2-3所示.

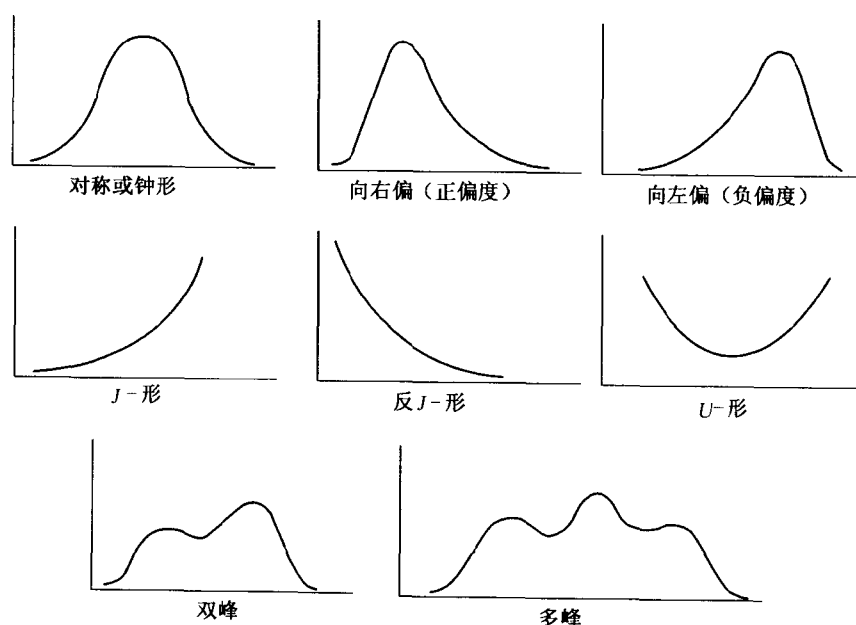


图 2-3

1. 当与中心最大值等距离的两边的观察值有相同的频数时,出现**对称的**或**钟形的**频数曲线.**正态曲线**是重要代表.

2. 在**微不对称的**或**斜的**频数曲线中,曲线一侧相对于中心最大值的尾部较另一侧的要长.如果长的尾部落在右边,这样的曲线称为**向右偏**或者说曲线有**正偏度**;反之,曲线称为**向左偏**或者说曲线有**负偏度**.

3. 在**J-形**或**反J-形**曲线中,最大值落在一端.

4. **U-形**频数曲线的两端都可取到极大值.

5. 一条**双峰**频数曲线有两个极大值.

6. 一条**多峰**频数曲线有多于两个的极大值.

习题及解答

数组阵列

2.1 (a) 排列数据 17, 45, 38, 27, 6, 48, 11, 57, 34 和 22.

(b) 求这些数的全距.

解 (a) 按数量大小升序排列为: 6, 11, 17, 22, 27, 34, 38, 45, 48, 57. 降序排列为: 57, 48, 45, 38, 34, 27, 22, 17, 11, 6.

(b) 由于最小值为 6, 最大值为 57, 所以全距为 $57 - 6 = 51$.

2.2 州立大学 80 个学生的数学学期成绩记录在下表中:

68	84	75	82	68	90	62	88	76	93
73	79	88	73	60	93	71	59	85	75
61	65	75	87	74	62	95	78	63	72
66	78	82	75	94	77	69	74	68	60
96	78	89	61	75	95	60	79	83	71
79	62	67	97	78	85	76	65	71	75
65	80	73	57	88	78	62	76	53	74
86	67	73	81	72	63	76	75	85	77

根据此表, 求:

- (a) 最高分;
- (b) 最低分;
- (c) 全距;
- (d) 前 5 名的成绩;
- (e) 后 5 名的成绩;
- (f) 第 10 名的成绩;
- (g) 成绩不低于 75 的学生人数;
- (h) 成绩低于 85 的学生人数;
- (i) 成绩介于 65 到 85 之间的学生所占的比例;
- (j) 没有出现的成绩.

解 其中的一些问题是很具体的, 因此我们最好先建立一个数组阵列. 把数据分为具体的组, 每个数据填入相应的组, 如表 2.3, 称为一个**登记表**. 然后对每一组数据进行排列, 如表 2.4, 就能得到我们需要的数组阵列. 根据表 2.4 我们就可以较容易地解答上述问题.

表 2.3

50~54	53
55~59	59, 57
60~64	62, 60, 61, 62, 63, 60, 61, 60, 62, 62, 63
65~69	68, 68, 65, 66, 69, 68, 67, 65, 65, 67
70~74	73, 73, 71, 74, 72, 74, 71, 71, 73, 74, 73, 72
75~79	75, 76, 79, 75, 75, 78, 78, 75, 77, 78, 75, 79, 79, 78, 76, 75, 78, 76, 76, 75, 77
80~84	84, 82, 82, 83, 80, 81
85~89	88, 88, 85, 87, 89, 85, 88, 86, 85
90~94	90, 93, 93, 94
95~99	95, 96, 95, 97

表 2.4

50~54	53
55~59	57, 59
60~64	60, 60, 60, 61, 61, 62, 62, 62, 62, 63, 63
65~69	65, 65, 65, 66, 67, 67, 68, 68, 68, 69
70~74	71, 71, 71, 72, 72, 73, 73, 73, 73, 74, 74, 74
75~79	75, 75, 75, 75, 75, 75, 75, 76, 76, 76, 76, 77, 77, 78, 78, 78, 78, 79, 79, 79
80~84	80, 81, 82, 82, 83, 84
85~89	85, 85, 85, 86, 87, 88, 88, 88, 89
90~94	90, 93, 93, 94
95~99	95, 95, 96, 97

- (a) 最高分是 97;
 (b) 最低分是 53;
 (c) 全距是 $97 - 53 = 44$;
 (d) 前 5 名的成绩为 97, 96, 95, 95 和 94;
 (e) 后 5 名的成绩为 53, 57, 59, 60 和 60;
 (f) 第 10 名的成绩为 88;
 (g) 成绩不低于 75 的学生人数为 44;
 (h) 成绩低于 85 的学生人数为 63;
 (i) 成绩介于 65 到 85 之间的学生所占的比例是 $49/80 = 61.2\%$;
 (j) 没有出现的成绩是 0~52, 54, 55, 56, 58, 64, 70, 91, 92, 98, 99 和 100.

频数分布, 直方图和频数多边形

2.3 表 2.5 给出了 P&R 公司 65 个员工周薪的频数分布. 根据表 2.5 求:

- (a) 第六组的下限;
 (b) 第四组的上限;
 (c) 第三组的组中值;
 (d) 第五组的组界;
 (e) 第五组组距的大小;
 (f) 第三组的频数;
 (g) 第三组的频率;
 (h) 频数最大的组距. 这通常称为**众数组**, 它的频数称为**众数组频数**;
 (i) 周薪低于 280.00 美元的员工比例;
 (j) 周薪介于 260.00~300.00 美元之间的员工比例.

表 2.5

工 资(美元)	职 工 数
250.00~259.99	8
260.00~269.99	10
270.00~279.99	16
280.00~289.99	14
290.00~299.99	10
300.00~309.99	5
310.00~319.99	2
总计	65

解 (a) 300.00 美元.

(b) 289.99 美元.

- (c) 第三组的组中值为 $\frac{1}{2} \times (270.00 + 279.99) = 274.995$. 实际应用中, 舍入至 275.00 美元.
- (d) 第五组的下组界为 $\frac{1}{2} \times (290.00 + 289.99) = 289.995$. 第五组的上组界为 $\frac{1}{2} \times (299.99 + 300.00) = 299.995$.
- (e) 第五组组距的大小 = 第五组的上组界 - 第五组的下组界 = $299.995 - 289.995 = 10.00$ 美元. 在此例中, 所有组距有相同的大小: 10.00 美元.
- (f) 16.
- (g) $16/65 = 0.246 = 24.6\%$.
- (h) 270.00 ~ 279.99.
- (i) 周薪低于 280.00 美元的职工人数是 $16 + 10 + 8 = 34$. 周薪低于 280.00 美元的职工比例为 $34/65 = 52.3\%$.
- (j) 周薪介于 260.00 ~ 300.00 美元之间的职工人数是 $10 + 14 + 16 + 10 = 50$. 周薪介于 260.00 ~ 300.00 美元之间的职工比例为 $50/65 = 76.9\%$.

2.4 如果在学生体重的频数分布中, 组中值分别为 128, 137, 146, 155, 164, 173 和 182 磅, 求: (a) 组距大小; (b) 组界; (c) 组限, 假定所有体重值均舍入至最近的磅数.

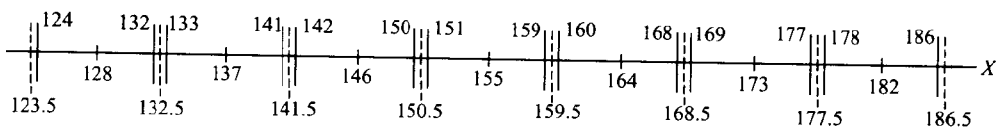
解 (a) 组距大小 = 连续组中值的共同差 = $137 - 128 = 146 - 137 = \dots = 9$ 磅.

(b) 由于组距有相同的大小, 因此组界是组中值间的中点, 值为 $\frac{1}{2} \times (128 + 137)$, $\frac{1}{2} \times (137 + 146)$, \dots , $\frac{1}{2} \times (173 + 182)$, 即 132.5, 141.5, 150.5, \dots , 177.5 磅. 由于共同的组距大小为 9 磅, 因此第一组的下组界是 $132.5 - 9 = 123.5$, 最后一组的上组界是 $177.5 + 9 = 186.5$. 所以所有的组界分别为: 123.5, 132.5, 141.5, 150.5, 159.5, 168.5, 177.5, 186.5 磅.

(c) 由于组限是整数, 我们选择最靠近组界的整数, 即, 123, 124, 132, 133, 141, 142, \dots . 因此第一组的组限为 124 ~ 132, 下一组的组限为 133 ~ 141, 以此类推.

2.5 用图表示习题 2.4 的结果.

解



如上图所示, 组中值 128, 137, 146, \dots , 182 均标在 X 轴上. 组界用长的垂直虚线表示, 组限由长的垂直实线表示.

2.6 150 个测量值的最小值为 5.18 英寸, 最大值为 7.44 英寸. 求一组可用于建立频数分布的恰当的 (a) 组距, (b) 组界, (c) 组中值.

解 全距为 $7.44 - 5.18 = 2.26$ 英寸. 如果定为 5 个组距, 组距大小约为 $2.26/5 = 0.45$; 如果定为 20 个组距, 组距大小大约为 $2.26/20 = 0.11$. 在 0.11 到 0.45 之间选择恰当的组距大小为 0.20, 0.30 或 0.40.

(a) 在下表中 I, II, III 分别表示组距大小为 0.20, 0.30 和 0.40 的恰当组距.

I	II	III
5.10 ~ 5.29	5.10 ~ 5.39	5.10 ~ 5.49
5.30 ~ 5.49	5.40 ~ 5.69	5.50 ~ 5.89
5.50 ~ 5.69	5.70 ~ 5.99	5.90 ~ 6.29
5.70 ~ 5.89	6.00 ~ 6.29	6.30 ~ 6.69
5.90 ~ 6.09	6.30 ~ 6.59	6.70 ~ 7.09
6.10 ~ 6.29	6.60 ~ 6.89	7.10 ~ 7.49
6.30 ~ 6.49	6.90 ~ 7.19	
6.50 ~ 6.69	7.20 ~ 7.49	
6.70 ~ 6.89		
6.90 ~ 7.09		
7.10 ~ 7.29		
7.30 ~ 7.49		

注意每一列第一组的下组限可以不为 5.10, 比如, 如果我们取 5.15 作为第 I 列第一组的下组限, 则第一组应写为 5.15~5.34.

(b) I, II, III 列对应的组界分别为

I	5.095~5.295, 5.295~5.495, 5.495~5.695, ..., 7.295~7.495
II	5.095~5.395, 5.395~5.695, 5.695~5.995, ..., 7.195~7.495
III	5.095~5.495, 5.495~5.895, 5.895~6.295, ..., 7.095~7.495

由于它们与测量值不一致, 因此这些组界是恰当的.

(c) I, II, III 列对应的组中值为

I	5.195, 5.395, ..., 7.395
II	5.245, 5.545, ..., 7.345
III	5.295, 5.695, ..., 7.295

这些组中值存在与测量值不一致的缺陷.

2.7 在解答习题 2.6(a) 时, 一个学生选择了组距 5.10~5.40, 5.40~5.70, ..., 6.90~7.20 和 7.20~7.50. 请问这样选择有问题吗?

解 这些组距在 5.40, 5.70, ..., 7.20 的点处出现重叠的现象. 因此一个测量值比如 5.40, 就可以放在前后两组的任何一组里. 一些统计学家通常选择把这些模棱两可的数据一半放在其中一组, 另一半放在另一组.

这种不确定的情况可以通过把组距写为 5.10 到低于 5.40, 5.40 到低于 5.70 等而得到改善. 此时, 组限与组界一致, 组中值与测量值一致.

通常地, 应尽可能避免重叠现象的发生, 并且组界应与实际观察值不一致. 例如, 为了避免习题 2.6 的组距产生不确定的情况, 组距应选为 5.095~5.395, 5.395~5.695, 等等. 但是, 这种特殊的选择会造成组中值与观察数据不一致.

2.8 下表记录了州立大学 40 个男同学的体重(舍入至最近的磅数), 建立频数分布.

138	164	150	132	144	125	149	157
146	158	140	147	136	148	152	144
168	126	138	176	163	119	154	165
146	173	142	147	135	153	140	135
161	145	135	142	150	156	145	128

解 最大体重值为 176 磅¹⁾, 最小体重值为 119 磅, 因此全距为 176~119=57 磅. 如果采用 5 个组距, 那么组距大小近似为 57/5=11; 如果采用 20 个组距, 那么组距大小近似为 57/20=3.

组距大小选择为 5 磅会显得方便些. 这样, 组中值可选择为 120, 125, 130, 135, ... 因此组距可选择为 118~122, 123~127, 128~132, ...; 组界为 117.5, 122.5, 127.5, ..., 这些值与观察值不一致.

所求的频数分布如表 2.6 所示. 中间一栏根据原始数据把组频数用唱票形式给出, 称为**计数或得分单**, 在最后的频数分布中得分单通常省略.

另解 当然, 存在着其他可能的频数分布. 例如, 表 2.7 给出了 7 个组, 组距为 9 磅的频数分布.

2.9 用 Minitab 建立习题 2.8 体重分布的(a)枝叶图, (b)直方图.

解 Minitab 软件产生的枝叶图如图 2-4(a)所示. 枝叶图由三列组成. 第二列表示给出数据的枝, 第三列表示这些数据的叶. 第一行 1, 11, 9 表示数据 119 的枝为 11, 叶为 9. 第二行 1, 12 表示数据 120 至 124 的枝为 12, 叶为 0, 1, 2, 3 或 4, 但体重数据中没有这几个数, 因此该行没有叶. 第三行 4, 12, 568 表示数据 125 至 129 的枝为 12, 叶为 5, 6, 7, 8 或 9. 体重数据表中只有 125, 126, 128 三个数, 因此这行的叶为 5, 6, 8. 余下各行类似. 图 2-4(a)的第一列在第七行的(8)以上数字表示从第一行开始的数据累计个数, 如第四行的 5 表示不大于 134 的数据共有 5 个; 第七行(8)以下的数字表示从倒数第一行开始

1) 1 磅 = 0.453592 公斤.

的数据累计个数,如倒数第三行的 4 表示不少于 165 的数据共有 4 个;第七行的(8)表示最先超过一半数据的那一行共有 8 个数据.

MTB > Stem-and-Leaf 'weight'.
Character Stem-and-Leaf Display

Stem-and-leaf of weight N = 40
Leaf Unit = 1.0

1 11 9
1 12
4 12 568
5 13 2
11 13 555688
17 14 002244
(8) 14 55667789
15 15 00234
10 15 678
7 16 134
4 16 58
2 17 3
1 17 6

(a)

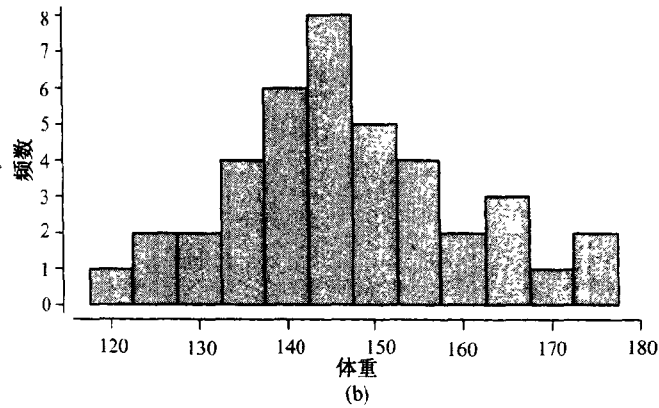


图 2-4

表 2.6

体重(磅)	计数	频数
118~122	一	1
123~127	丁	2
128~132	丁	2
133~137	正	4
138~142	正一	6
143~147	正下	8
148~152	正	5
153~157	正	4
158~162	丁	2
163~167	下	3
168~172	一	1
173~177	丁	2
总计		40

表 2.7

体重(磅)	计数	频数
118~126	下	3
127~135	正	5
136~144	正正	9
145~153	正正丁	12
154~162	正	5
163~171	正	4
172~180	丁	2
总计		40

2.10 根据习题 2.3 表 2.5, 建立(a)频率分布;(b)频数直方图;(c)频率直方图;(d)频数多边形;(e)频率多边形.

解 (a)在表 2.5 的频数分布中,用每一组的组频数除以总频数(65),得到频率分布,结果用百分数表示,如表 2.8 所示.
(b)和(c)频数直方图和频率直方图如图 2-5 所示.注意,把一个频数直方图转化为一个频率直方图只要加一个竖直标度表示频率,如图 2-5 的右侧.

表 2.8

工资(元)	频率(百分数表示)
250.00~259.99	12.3
260.00~269.99	15.4
270.00~279.99	24.6
280.00~289.99	21.5
290.00~299.99	15.4
300.00~309.99	7.7
310.00~319.99	3.1
	总计 100.0

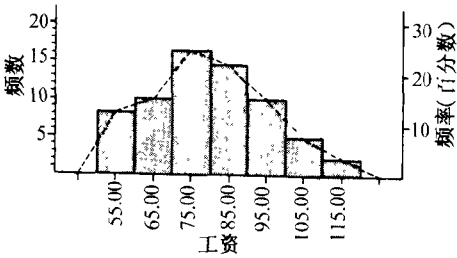


图 2-5

(d)和(e)在图 2-5 中用虚线表示频数多边形和频率多边形.把一个频数多边形转化为一个频率多边形,只需要加一个竖直标度表示频率.

注意,如果只需要频率多边形,那么图形可不包含直方图,频率轴应放置在左边代替频数轴.

2.11 证明直方图中矩形总面积等于相应的频数多边形和 X 轴围成区域的总面积.

证明 假定直方图由 3 个矩形构成,如图 2-6 所示,其中虚线表示相应的频数多边形.

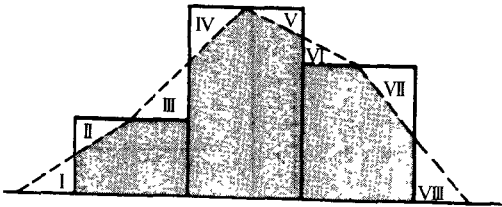


图 2-6

$$\begin{aligned}
 \text{矩形总面积} &= \text{阴影面积} + \text{II 的面积} + \text{IV 的面积} + \text{V 的面积} + \text{VII 的面积} \\
 &= \text{阴影面积} + \text{I 的面积} + \text{III 的面积} + \text{VI 的面积} + \text{VIII 的面积} \\
 &= \text{频数多边形和 X 轴围成的区域总面积}
 \end{aligned}$$

其中我们知道 I 的面积 = II 的面积, III 的面积 = IV 的面积, V 的面积 = VI 的面积, VII 的面积 = VIII 的面积.

2.12 在 P&R 公司(见习题 2.3),5 个新员工的周薪是 285.34 美元,316.83 美元,335.78 美元,356.21 美元和 374.50 美元.对这新老 70 个职员建立频数分布.

解 可能的频数分布如表 2.9 所示.

在表 2.9(a)中,相等的组距大小为 10.00 美元.因此,有许多空组并且在工资标度的尾部细节太多.

在表 2.9(b)中,空组和过多的细节通过开组距“320.00 美元以上”的使用而避免.这样做的不利之处在于完成某些数学计算时,表格会失去效用.例如,由于“320.00 美元以上”可能表示了个人每周挣得 1400.00 美元,因此就不可能确定每周支付的工资总数.

在表 2.9(c)中,组距大小为 20.00 美元.它的不利之处在于在工资标度的前面部分有许多信息被忽略掉了,并且在工资标度的尾部细节仍然较多.

在表 2.9(d)中,组距大小不相等.它的不利之处在于在以后的某些数学计算中,不如组距大小相等时那么简便.组距愈大,产生的误差也会愈多.

表 2.9(a)

工资(美元)	频数
250.00~259.99	8
260.00~269.99	10
270.00~279.99	16
280.00~289.99	15
290.00~299.99	10
300.00~309.99	5
310.00~319.99	3
320.00~329.99	0
330.00~339.99	1
340.00~349.99	0
350.00~359.99	1
360.00~369.99	0
370.00~379.99	1
总计	70

表 2.9(b)

工资(美元)	频数
250.00~259.99	8
260.00~269.99	10
270.00~279.99	16
280.00~289.99	15
290.00~299.99	10
300.00~309.99	5
310.00~319.99	3
320.00 以上	3
总计	70

表 2.9(c)

工资(美元)	频数
250.00~269.99	18
270.00~289.99	31
290.00~309.99	15
310.00~329.99	3
330.00~349.99	1
350.00~369.99	1
370.00~389.99	1
总计	70

表 2.9(d)

工资(美元)	频数
250.00~259.99	8
260.00~269.99	10
270.00~279.99	16
280.00~289.99	15
290.00~299.99	10
300.00~319.99	8
320.00~379.99	3
总计	70

2.13 根据表 2.9(d)的频数分布建立直方图.

解 所求直方图如图 2-7 所示.为了建立此直方图,我们采用面积与频数成比例的方法.假定矩形 A 对应第一组(见表 2.9(d)),组频数为 8.由于表 2.9(d)的第六组也有频数 8,代表这个组的矩形 B 也应该有着与 A 相同的面积.由于 B 宽是 A 宽的两倍,因此它的高应是 A 高的一半.同理,代表表 2.9(d)最后一组的矩形 C,在竖直标度上应是半个单位高.

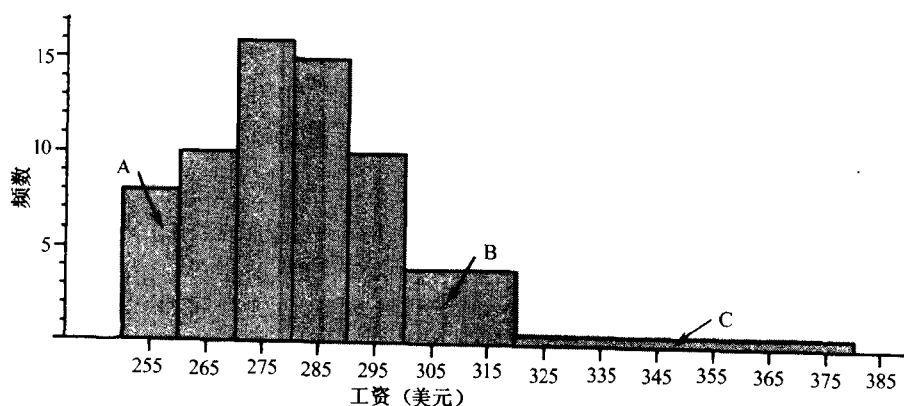


图 2-7

累积频数分布和卵形线

2.14 根据习题 2.3 表 2.5 的频数分布作:(a)累积频数分布;(b)累积频率分布;(c)卵形线;(d)百分率卵形线.

解 (a)和(b)累积频数分布和累积频率分布如表 2.10 所示.

注意第二列每行的值是把表 2.5 的第二列相应行及以上行的值加起来而得到的,比如 $18 = 8 + 10$, $34 = 8 + 10 + 16$, 等等.

第三列的每一个值是把前一列的值除以总频数 65 而得到的,结果表示成百分数.比如, $34/65 = 52.3\%$.这一列各行的值也可由表 2.8 的第二列相应行及以上行的值加起来得到.因此, $27.7 = 12.3 + 15.4$, $52.3 = 12.3 + 15.4 + 24.6$, 等等.

表 2.10

工资(美元)	累积频数	累积频率(百分数)
低于 250.00	0	0.0
低于 260.00	8	12.3
低于 270.00	18	27.7
低于 280.00	34	52.3
低于 290.00	48	73.8
低于 300.00	58	89.2
低于 310.00	63	96.9
低于 320.00	65	100.0

(c)和(d)卵形线(或累积频数多边形)和百分率卵形线分别如图 2-8(a)和(b)所示.它们均是由 Minitab 产生的.

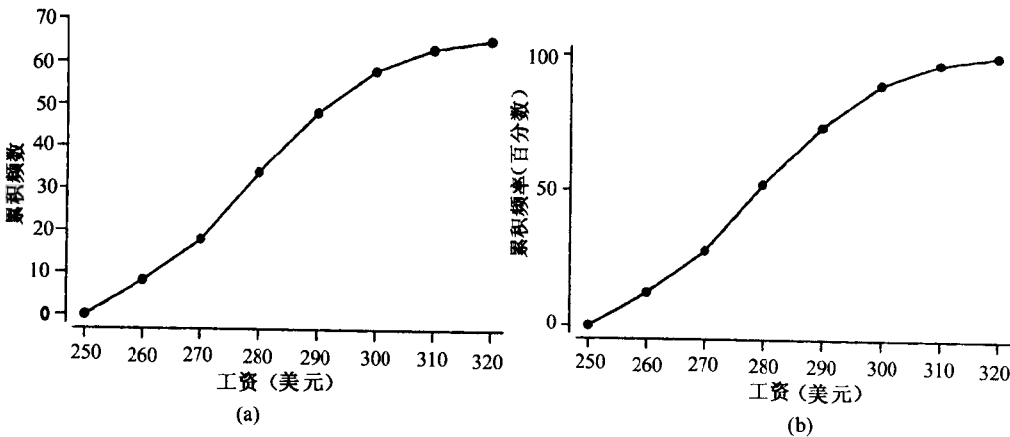


图 2-8

2.15 根据习题 2.3 表 2.5 的频数分布,建立:(a)“不低于”累积频数分布;(b)“不低于”卵形线.

解 (a)表 2.11 第二列每行的值都是表 2.5 第二列相应行及以下行的值的和,开始于表 2.5 的底部,比如, $7 = 2 + 5$, $17 = 2 + 5 + 10$, 等等.这些值也可用总频数 65 减去表 2.10 第二列相应行的值而得到,比如, $57 = 65 - 8$, $47 = 65 - 18$, 等等.

(b)图 2-9 给出了“不低于”卵形线.

表 2.11

工资(美元)	“不低于”累积频数
不低于 250.00	65
不低于 260.00	57
不低于 270.00	47
不低于 280.00	31
不低于 290.00	17
不低于 300.00	7
不低于 310.00	2
不低于 320.00	0

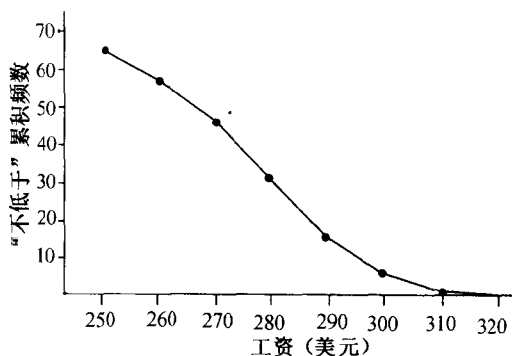


图 2-9

- 2.16 根据图 2-8 和图 2-9(分别关于习题 2.14 和 2.15)的卵形线,估计满足条件的职工人数:(a)周薪低于 288.00 美元,(b)周薪不低于 296.00 美元,(c)周薪介于 263.00 到 275.00 美元之间.

解 (a)根据图 2-8 的“低于”卵形线,作一条竖直线交“工资”轴于 288.00 美元.这条线交卵形线于点(288,45),因此,45 个职工周薪低于 288.00 美元.

(b)根据图 2-9 的“不低于”卵形线,作一条竖直线交“工资”轴于 296.00 美元.这条线交卵形线于点(296,11),因此,11 个职工周薪不低于 296.00 美元.

此结果也可通过图 2-8 的“低于”卵形线得到.在 296.00 美元处作竖直线,我们发现 54 个职工周薪低于 296.00 美元,因此 $65 - 54 = 11$ 个职工周薪不低于 296.00 美元.

(c)根据图 2-8 的“低于”卵形线,得到:所求职工数 = 周薪低于 275.00 美元人数 - 周薪低于 263.00 美元人数 = $26 - 11 = 15$.

上述结果也可通过累积频数表的线性插值得到.例如,在(a)中,因为 $288 = 280 + 8 = 280 + \frac{8}{10} \times (290 - 280)$,而 280 和 290 相对应的累积频数分别为 34 和 48(见表 2.10), $(48 - 34) \times \frac{8}{10} = 11$,因此所求人数为 $34 + 11 = 45$.

- 2.17 抛掷 5 枚硬币 1000 次,每次抛掷都记录下正面次数.表 2.12 记录了正面次数为 0,1,2,3,4 和 5 的抛掷次数.

(a) 根据表 2.12 的数据作图.

表 2.12

正面次数	抛掷次数(频数)
0	38
1	144
2	342
3	287
4	164
5	25
总计	1000

(b) 建一个表格来表示正面次数低于 0,1,2,3,4,5 或 6 的抛掷次数的百分比.

(c) 根据(b)中表的数据作图.

解 (a) 可用图 2-10 或 2-11 表示 2.12 中数据的情况.

由于正面数不能为 1.5 或 3.2,因此图 2-10 看上去更自然.此图是条形图的一种,此时条带宽是 0,有时也把这样的图称为杆图.遇到离散数据时,多采用此图形.

图 2-11 是数据的直方图.注意,直方图的总面积应是总频数 1000.在作直方图或相应的频数多边形时,我们通常假定数据是连续的.在后继学习中,这一点将会有用的.注意,我们在习题 2.10 中已对离散数据作过直方图或频数多边形.

(b) 所求作的表如表 2.13 所示,这里只显示正面次数的累积频数分布和百分率频数分布.术语“少于 1”,“少于 2”等等应理解为“少于或等于 0”,“少于或等于 1”等等.

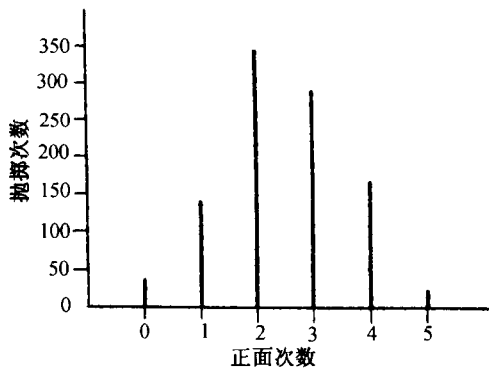


图 2-10

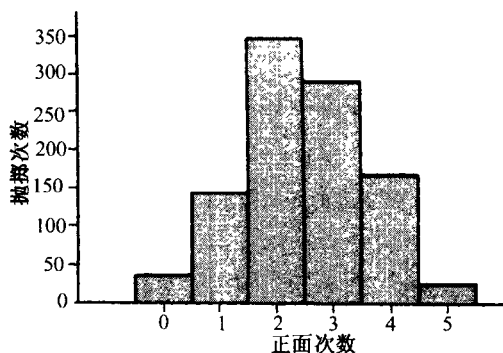


图 2-11

表 2.13

正面次数	抛掷次数 (累积频数)	抛掷次数百分率 (百分率累积频数)
少于 0	0	0.0
少于 1	38	3.8
少于 2	182	18.2
少于 3	524	52.4
少于 4	811	81.1
少于 5	975	97.5
少于 6	1000	100.0

(c)所求作图形,如图 2-12 或 2-13 所示.

由于正面次数少于 2 的抛掷次数所占百分率等于正面次数少于 1.75, 1.56 或 1.23 的抛掷所占百分率,因此对这些值来说,百分率都等于 18.2% (用水平线表示). 因此,图 2-12 用来显示离散数据的情况是很恰当的.

图 2-13 显示的是数据的累积频数多边形或卵形线,这里也把数据看做连续的.

图 2-12 和 2-13 分别对应(a)中的图 2-10 和 2-11.

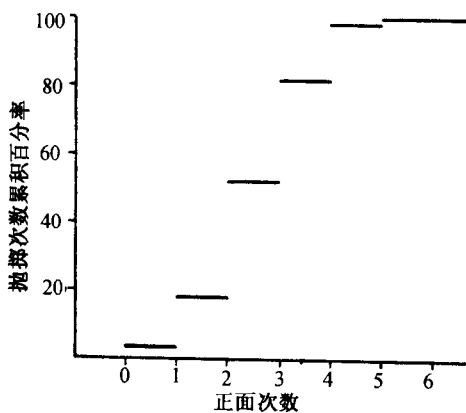


图 2-12

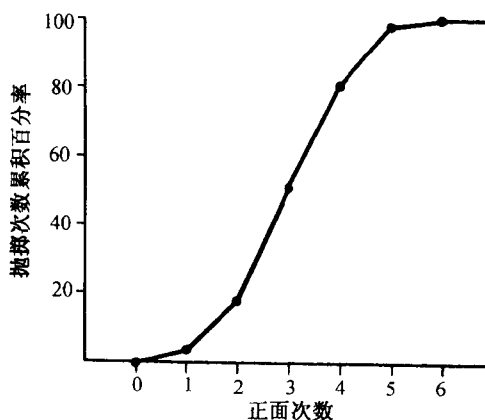


图 2-13

频数曲线和光滑卵形线

2.18 从 XYZ 大学 1546 个男同学中抽取 100 个作为一个样本(表 2.1).

(a)根据样本中提供的数据,建立一个光滑百分率频数多边形(频率曲线)和一条光滑的

“低于”百分率卵形线。

(b)根据(a)中结论,估计该大学中身高在 65~70 英寸之间的学生数.解决这个问题,你需要哪些假定?

(c)你能根据上述结论估计美国男学生身高在 65~70 英寸之间的比例吗?

解 (a)图 2-14 和 2-15 中虚线分别代表的是从图 2-1 和 2-2 中得到的频率多边形和百分率卵形线.所求作的光滑图形(黑线表示)是用光滑曲线近似它们而得.

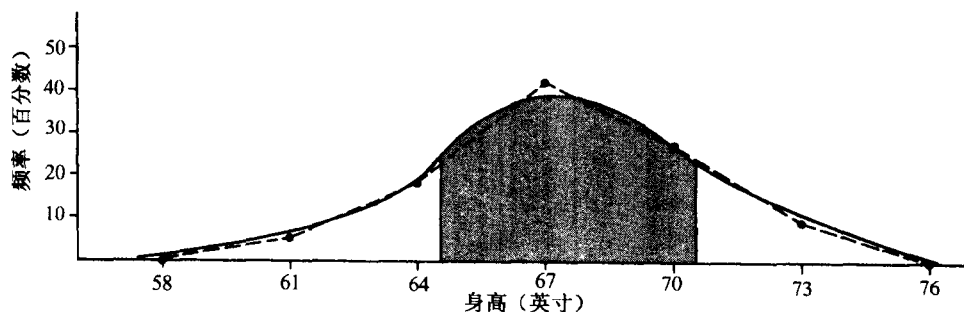


图 2-14

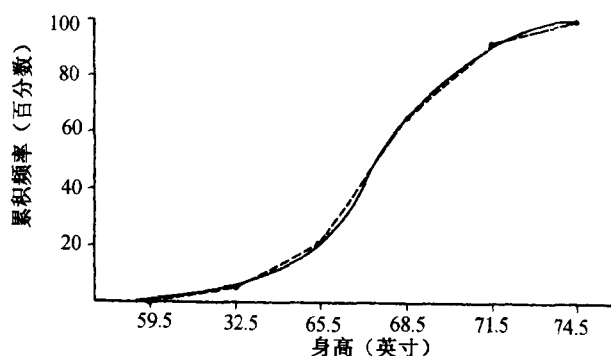


图 2-15

在实际应用中,因为较容易得到光滑百分率卵形线,所以先作光滑百分率卵形线,再从其上读取数值得到光滑频率多边形。

(b)如果 100 个学生的样本是总体为 1546 个学生的代表,那么图 2-14 和 2-15 的光滑曲线可假定是总体的频率曲线和光滑百分率卵形线.这个假定仅当样本是**随机抽取**(即每个同学被选择的机会是相等的)时是正确的。

由于身高在 65~70 英寸之间实际上代表的是身高在 64.5~70.5 英寸之间,因此总体中对应身高的学生所占的百分比可通过图 2-14 中阴影部分面积除以光滑曲线和 X 轴围成的面积而得。

用图 2-15 会更简便.从图中我们看出:

身高低于 70.5 英寸的学生所占百分比 = 82 %

身高低于 64.5 英寸的学生所占百分比 = 18 %

所以身高在 64.5~70.5 英寸之间的学生所占百分比 = 82 % - 18 % = 64 % . 现在可知,在 XYZ 大学身高在 65~70 英寸之间的学生数是 $1546 \times 64 \% = 989$ 人。

这也可由**概率**来描述,即从 1546 人中随机抽取的学生身高在 65~70 英寸之间的概率是 64 % 或 0.64. 由于与概率的联系(详见第六章),相对频率曲线也常称为**概率曲线**或**概率分布**。

(c)仅当我们确信抽取的 100 个学生的样本是从美国所有男学生总体中随机抽取时,才能认为所求比例是 64 % (因此,这个比例比(b)有更多的不确定性)。然而,由于某些原因,这看上去也许不可能,比如:(1)一些大学学生仍然在长个,(2)年轻一代比他们的父辈要高等等。

补充习题

2.19 (a)排列数据 12, 56, 42, 21, 5, 18, 10, 3, 61, 34, 65 和 24; (b)求这些数据的全距.

2.20 表 2.14 是 400 个初中生每周看电视时间(以分钟计)的频数分布. 根据此表计算:

(a)第五组上组限;

(b)第八组下组限;

(c)第七组组中值;

(d)最后一组组界;

(e)组距大小;

(f)第四组频数;

(g)第六组的频率;

(h)每周看电视时间不超过 600 分钟的学生所占百分比;

(i)每周看电视时间多于或等于 900 分钟的学生所占百分比;

(j)每周看电视时间在 500 ~ 1000 分钟之间的学生所占百分比.

表 2.14

观看时间(分钟)	学生数
300 ~ 399	14
400 ~ 499	46
500 ~ 599	58
600 ~ 699	76
700 ~ 799	68
800 ~ 899	62
900 ~ 999	48
1000 ~ 1099	22
1100 ~ 1199	6

2.21 根据表 2.14 的频数分布建立: (a)直方图; (b)频数多边形.

2.22 根据习题 2.20 表 2.14 建立: (a)频率分布; (b)频率直方图; (c)频率多边形.

2.23 根据表 2.14 中数据建立: (a)累积频数分布; (b)百分率累积分布; (c)卵形线; (d)百分率卵形线(除非特别说明, 累积频数是在“低于”基础上).

2.24 累积频数在“不低于”基础上解答习题 2.23.

2.25 根据表 2.14 中数据, 估计每周看电视: (a)少于 560 分钟; (b)多于或等于 970 分钟; (c)在 620 ~ 890 分钟之间的学生数.

2.26 一公司生产的垫圈内直径测量至最近的千分之一英寸. 如果这些直径频数分布的组中值为 0.321, 0.324, 0.327, 0.330, 0.333, 0.336, 求(a)组距大小; (b)组界; (c)组限.

2.27 下表给出一公司生产的 60 个抽样轴承滚珠直径(以厘米计). 选择适当的组距, 建立直径的频数分布.

1.738	1.729	1.743	1.740	1.736	1.741	1.735	1.731	1.726	1.737
1.728	1.737	1.736	1.735	1.724	1.733	1.742	1.736	1.739	1.735
1.745	1.736	1.742	1.740	1.728	1.738	1.725	1.733	1.734	1.732
1.733	1.730	1.732	1.730	1.739	1.734	1.738	1.739	1.727	1.735
1.735	1.732	1.735	1.727	1.734	1.732	1.736	1.741	1.736	1.744
1.732	1.737	1.731	1.746	1.735	1.735	1.729	1.734	1.730	1.740

2.28 根据习题 2.27 中数据建立: (a)直方图; (b)频数多边形; (c)频率分布; (d)频率直方图; (e)频率多边形; (f)累积频数分布; (g)百分率累积分布; (h)卵形线; (i)百分率卵形线.

2.29 根据习题 2.28 中结论, 求满足下列条件的滚珠轴承直径的百分比: (a)超过 1.732 cm; (b)不多于 1.736 cm; (c)在 1.730 ~ 1.738 cm 之间. 把这里的结论与从习题 2.27 中原始数据得到的结论作比较.

2.30 利用习题 2.20 中数据完成习题 2.28.

2.31 根据美国人口调查局最新人口报告, 1996 年美国人口数是 265 284 000. 表 2.15 给出不同年龄组的百分率分布:

(a)第二组的组距大小是多少? 第四组呢?

(b)有多少种不同的组距大小?

(c)有多少个开组距?

(d)应如何确定最后一组才能使它的组距与前一组组距大小相等?

(e)第二组组中值是多少? 第四组呢?

(f)第四组的组界是多少?

(g)年龄大于等于 35 岁人口所占比例是多大? 年龄小于等于 64 岁呢?

(h)年龄在 20 ~ 49 岁之间人口所占比例是多大?

(i) 年龄大于 70 岁人口所占比例是多大?

表 2.15

年龄组	百分率
5 以下	7.3
5~9	7.3
10~14	7.2
15~19	7.0
20~24	6.6
25~29	7.2
30~34	8.1
35~39	8.5
40~44	7.8
45~49	6.9
50~54	5.3
55~59	4.3
60~64	3.8
65~74	7.0
75~84	4.3
85 及 85 以上	1.4

来源:美国人口调查局最新人口报告

- 2.32 (a) 为什么根据表 2.15 中分布不能建立百分率直方图或百分率频数多边形?
 (b) 要建立百分率直方图或百分率频数多边形需对现有分布作怎样的修改?
 (c) 利用(b)中修改完成作图.
- 2.33 在表 2.15 中, 假设总人口为 265 000 000, 并且“小于 5”这一组包括不到 1 岁的婴儿. 确定每组人数(以百万计, 并保留一位小数).
- 2.34 (a) 根据表 2.14 中数据建立光滑百分率频数多边形和光滑百分率卵形线.
 (b) 利用(a)中结论, 估计每周看电视少于 10 小时的学生的概率.
 (c) 利用(a)中结论, 估计每周看电视等于或多于 15 小时的学生的概率.
 (d) 利用(a)中结论, 估计每周看电视少于 5 小时的学生的概率.
- 2.35 (a) 抛掷四枚硬币 50 次, 建立每次抛掷正面数表.
 (b) 建立频数分布来显示正面次数是 0, 1, 2, 3, 4 的抛掷数.
 (c) 根据(b)建立相应的百分率分布.
 (d) 把(c)中得到的百分率与理论上的 6.25%, 25%, 37.5%, 25%, 6.25% (与 1, 4, 6, 4, 1 成比例) 作比较.
 (e) 根据(a)和(b)中分布作图.
 (f) 建立数据的百分率卵形线.
- 2.36 抛掷四枚硬币多于 50 次, 看看这样是不是与理论值更一致? 如果不是, 请给出可能产生差异的原因.

第三章 均值, 中位数, 众数以及其他 表示集中趋势的度量

下标, 记法

符号 X_j 表示变量 X 的 N 个值 $X_1, X_2, X_3, \dots, X_N$ 中的任意一个. 在 X_j 中, 可代表数字 $1, 2, 3, \dots, N$ 中任何一个的字母 j , 称为下标. 显然, 其他的一些字母, 比如 i, k, p, q 或 s 都可以作为下标.

求和符号

符号 $\sum_{j=1}^N X_j$ 表示所有从 $j=1$ 到 $j=N$ 的 X_j 之和. 由定义可知

$$\sum_{j=1}^N X_j = X_1 + X_2 + X_3 + \dots + X_N$$

在不产生混淆的情况下, 我们常简单的记为 $\sum X, \sum X_j, \sum_j X_j$. 符号 \sum 是 σ 的大写希腊字母, 表示求和.

例 1 $\sum_{j=1}^N X_j Y_j = X_1 Y_1 + X_2 Y_2 + X_3 Y_3 + \dots + X_N Y_N.$

例 2 $\sum_{j=1}^N aX_j = aX_1 + aX_2 + aX_3 + \dots + aX_N = a(X_1 + X_2 + X_3 + \dots + X_N) = a \sum_{j=1}^N X_j,$

其中 a 是一个常数. 简单地表示为: $\sum aX = a \sum X.$

例 3 若 a, b 和 c 是任意常数, 则 $\sum (aX + bY - cZ) = a \sum X + b \sum Y - c \sum Z$. 参见习题 3.3.

平均值或集中趋势的度量

平均值是一组数据典型的或有代表性的值. 由于这样的典型值趋向于落在根据数值大小排列的数据的中心, 因此平均值也称为**集中趋势的度量**.

可以定义几种类型的平均值, 最常用的有**算术平均**, **中位数**, **众数**, **几何平均**及**调和平均**. 根据数据情况和使用的目的, 每一类平均值都各有利弊.

算术平均

N 个数 $X_1, X_2, X_3, \dots, X_N$ 的**算术平均**或简称**均值**用 \bar{X} 表示, 定义为

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \dots + X_N}{N} = \frac{\sum_{j=1}^N X_j}{N} = \frac{\sum X}{N} \quad (1)$$

例 4 8, 3, 5, 12 和 10 的算术平均值为

$$\bar{X} = \frac{8 + 3 + 5 + 12 + 10}{5} = \frac{38}{5} = 7.6$$

如果 X_1, X_2, \dots, X_K 分别出现 f_1, f_2, \dots, f_K 次 (即以频数 f_1, f_2, \dots, f_K 出现), 那么算术平均值为

$$\bar{X} = \frac{f_1 X_1 + f_2 X_2 + \dots + f_K X_K}{f_1 + f_2 + \dots + f_K} = \frac{\sum_{j=1}^K f_j X_j}{\sum_{j=1}^K f_j} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N} \quad (2)$$

其中 $N = \sum f$ 是**总频数**(即数字出现的总次数).

例 5 如果 5, 8, 6 和 2 分别出现 3, 2, 4 和 1 次, 那么它们的算术平均值为

$$\bar{X} = \frac{3 \times 5 + 2 \times 8 + 4 \times 6 + 1 \times 2}{3 + 2 + 4 + 1} = \frac{15 + 16 + 24 + 2}{10} = 5.7$$

加权算术平均

有时, 根据各个数字的显著性和重要性, 我们需要在 X_1, X_2, \dots, X_K 上加某些**加权因子**(或权) w_1, w_2, \dots, w_K . 此时

$$\bar{X} = \frac{w_1 X_1 + w_2 X_2 + \dots + w_K X_K}{w_1 + w_2 + \dots + w_K} = \frac{\sum wX}{\sum w} \quad (3)$$

称为**加权算术平均**. 注意, (2)式与(3)式有相似性, 因此(2)式可视为权是 f_1, f_2, \dots, f_K 的加权算术平均.

例 6 设一门功课期末考试成绩的权是小测验成绩权的 3 倍, 若一个学生的期末考试成绩为 85 分, 小测验成绩为 70 和 90 分, 则平均分为

$$\bar{X} = \frac{1 \times 70 + 1 \times 90 + 3 \times 85}{1 + 1 + 3} = \frac{415}{5} = 83$$

算术平均的性质

1. 一组数与它们的算术平均之差的代数和为零.

例 7 8, 3, 5, 12 和 10 与它们的算术平均 7.6 的差分别为 $8 - 7.6, 3 - 7.6, 5 - 7.6, 12 - 7.6$ 和 $10 - 7.6$, 即 0.4, -4.6, -2.6, 4.4 和 2.4, 这些差的代数和为 $0.4 - 4.6 - 2.6 + 4.4 + 2.4 = 0$.

2. 一组数 X_j 与任意数 a 之差的平方和当且仅当 $a = \bar{X}$ 时达到最小(见习题 4.27).

3. 如果 f_1 个数有平均值 m_1, f_2 个数有平均值 m_2, \dots, f_K 个数有平均值 m_K , 那么这些数的平均值为

$$\bar{X} = \frac{f_1 m_1 + f_2 m_2 + \dots + f_K m_K}{f_1 + f_2 + \dots + f_K} \quad (4)$$

即所有平均值的加权平均(见习题 3.12).

4. 如果 A 是任一**假定算术平均值**(可以为任何数), 并记 $d_j = X_j - A$, 那么(1)和(2)分别变为

$$\bar{X} = A + \frac{\sum_{j=1}^N d_j}{N} = A + \frac{\sum d}{N} \quad (5)$$

$$\bar{X} = A + \frac{\sum_{j=1}^K f_j d_j}{\sum_{j=1}^K f_j} = A + \frac{\sum fd}{N} \quad (6)$$

其中 $N = \sum_{j=1}^K f_j = \sum f$. 注意, (5)和(6)式实际上可概括为 $\bar{X} = A + \bar{d}$ (见习题 3.18).

从分类资料中计算算术平均值

当数据以频数分布的形式表现, 我们视所有落在某一给定组距里的值与组中值一致. 如果我们认为 X_j 是第 j 个组距的组中值, f_j 是相应的组频数, A 是一假定组中值, 且 $d_j = X_j - A$ 表示 X_j 与 A 的差, 那么(2)和(6)式可用于分类资料.

使用(2)和(6)式的计算分别称为**长**和**短方法**(见习题 3.15 和 3.20).

如果组距都有相等的宽度 c , 那么 $d_j = X_j - A$ 可表为 cu_j , 其中 u_j 可是正、负整数或零

(即, $0, \pm 1, \pm 2, \pm 3, \dots$), 并且(6)式可变为

$$\bar{X} = A + \frac{\sum_{j=1}^K f_j c u_j}{N} = A + \left(\frac{\sum f u}{N} \right) c \quad (7)$$

此式与等式 $\bar{X} = A + c \bar{u}$ 是等价的(见习题 3.21). 这种方法称为计算平均值的**编码法**. 这是一种短方法, 可以用来对组距宽度相等时的分类资料进行计算(见习题 3.22 和 3.23). 在编码法中, 变量 X 的值根据 $X = A + cu$ **变换**为变量 u 的值.

中位数

一组数根据数量大小排列后的中间值或者两个中间值的算术平均值称为这组数的**中位数**.

例 8 一组数 3, 4, 4, 5, 6, 8, 8, 8 和 10 的中位数是 6.

例 9 一组数 5, 5, 7, 9, 11, 12, 15 和 18 的中位数是 $\frac{1}{2}(9 + 11) = 10$.

对于分类资料, 用插值法求中位数的公式为

$$\text{中位数} = L_1 + \left(\frac{\frac{N}{2} - (\sum f)_1}{f_m} \right) c \quad (8)$$

其中

L_1 = 中位数组(即包含中位数的组)的下组界

N = 数据的总个数

$(\sum f)_1$ = 中位数组前各组的频数和

f_m = 中位数组频数

c = 中位数组组距宽度

从几何上看, 中位数是分直方图为二等份面积的垂直线的 X (横坐标)的值. 这个 X 的值常用 \tilde{X} 表示.

众数

一组数的**众数**是出现次数最多的那个数, 即以最大频数出现的数. 众数不一定存在, 即使存在也不必惟一.

例 10 一组数 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12 和 18 的众数为 9.

例 11 一组数 3, 5, 8, 10, 12, 15 和 16 无众数.

例 12 一组数 2, 3, 4, 4, 4, 5, 5, 7, 7, 7 和 9 有两个众数: 4 和 7, 称为**双峰**.

只有一个众数的分布称为**单峰**的.

当分类资料的频数曲线用来拟合数据时, 众数是曲线上最大值点的 X 的值. 这个 X 的值也常用 \hat{X} 表示.

根据频数分布或直方图计算众数的公式为

$$\text{众数} = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) c \quad (9)$$

其中,

L_1 = 众数组(即包含众数的组)的下组界.

Δ_1 = 众数频数减去前一组的频数.

Δ_2 = 众数频数减去后一组的频数.

c = 众数组组距宽度.

均值, 中位数和众数间的经验关系

对于微斜(不对称)的单峰频数曲线, 我们有经验关系式:

$$\text{均值} - \text{众数} = 3(\text{均值} - \text{中位数}) \quad (10)$$

图 3-1 和图 3-2 分别表示了向右、向左倾斜的频数曲线的均值, 中位数和众数的相对位置. 对于对称曲线, 均值、中位数和众数是完全一致的.

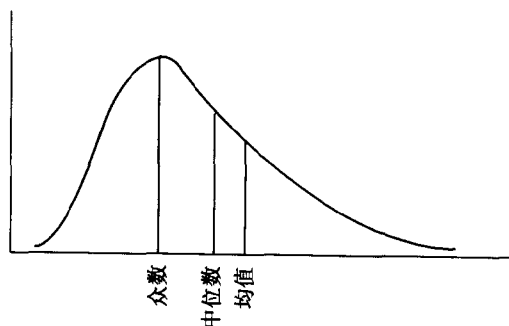


图 3-1

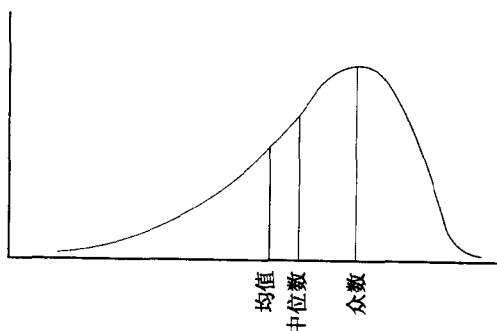


图 3-2

几何平均 G

N 个正数 X_1, X_2, \dots, X_N 的几何平均 G 等于这些数乘积的 N 次方根:

$$G = \sqrt[N]{X_1 X_2 X_3 \cdots X_N} \quad (11)$$

例 13 数 2, 4 和 8 的几何平均值为 $G = \sqrt[3]{2 \times 4 \times 8} = \sqrt[3]{64} = 4$.

我们可以用对数法(见习题 3.35)或计算器来求解 G . 对于如何从分类资料中求解几何平均的问题, 参看习题 3.36 和 3.91.

调和平均 H

N 个数 X_1, X_2, \dots, X_N 的调和平均 H 等于这些数的倒数的算术平均的倒数:

$$H = \frac{1}{\frac{1}{N} \sum_{j=1}^N \frac{1}{X_j}} = \frac{N}{\sum \frac{1}{X}} \quad (12)$$

实际应用中, 简单记为

$$\frac{1}{H} = \frac{\sum \frac{1}{X}}{N} = \frac{1}{N} \sum \frac{1}{X} \quad (13)$$

例 14 数 2, 4 和 8 的调和平均值为

$$H = \frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{8}} = \frac{3}{\frac{7}{8}} = 3.43$$

对于如何从分类资料中求解调和平均的问题, 参看习题 3.99 和 3.100.

算术平均, 几何平均和调和平均间的关系

一组正数 X_1, X_2, \dots, X_N 的几何平均小于等于它们的算术平均, 但大于等于它们的调和平均. 用符号表示为

$$H \leq G \leq \bar{X} \quad (14)$$

当所有的数 X_1, X_2, \dots, X_N 相等时, 等号成立.

例 15 2, 4, 8 的算术平均是 4.67, 几何平均是 4, 调和平均是 3.43.

均方根(RMS)

一组数 X_1, X_2, \dots, X_N 的均方根或二次平均常用 $\sqrt{X^2}$ 表示, 并定义为

$$\text{均方根} = \sqrt{X^2} = \sqrt{\frac{\sum_{j=1}^N X_j^2}{N}} = \sqrt{\frac{\sum X^2}{N}} \quad (15)$$

这种类型的平均常在实际中使用.

例 16 1, 3, 4, 5 和 7 的均方根为

$$\sqrt{\frac{1^2 + 3^2 + 4^2 + 5^2 + 7^2}{5}} = \sqrt{20} = 4.47$$

四分位数, 十分位数和百分位数

一组数据按照数量大小排列, 如果中间的数(或两个中间数的算术平均)把这组数分为 2 个相等部分, 那么这样的数称为中位数. 按照这种思路, 我们想到了那些把一组数分为 4 个相等部分的数. 这些数用 Q_1, Q_2 和 Q_3 表示, 分别称为第一, 第二和第三个四分位数, 其中 Q_2 等于中位数.

同样地, 把一组数分为 10 个相等部分的数称为十分位数, 并且用 D_1, D_2, \dots, D_9 表示, 而把一组数分为 100 个相等部分的数称为百分位数, 用 P_1, P_2, \dots, P_{99} 表示. 其中第五个十分位数和第 50 个百分位数与相应的中位数一致, 第二十五个和第七十五个百分位数分别与第一个和第三个四分位数相等.


四分位数, 十分位数, 百分位数及其他一些通过等分数据而得到的数统称为分位数. 从分类资料中计算这些值, 参见习题 3.44 到 3.46.

习题及解答

求和符号


3.1 写出下列各式的展开式:

$$\begin{aligned} \text{(a)} & \sum_{j=1}^6 X_j; & \text{(c)} & \sum_{j=1}^N a; & \text{(e)} & \sum_{j=1}^3 (X_j - a). \\ \text{(b)} & \sum_{j=1}^4 (Y_j - 3)^2; & \text{(d)} & \sum_{k=1}^5 f_k X_k; \end{aligned}$$

解  (a) $X_1 + X_2 + X_3 + X_4 + X_5 + X_6$
(b) $(Y_1 - 3)^2 + (Y_2 - 3)^2 + (Y_3 - 3)^2 + (Y_4 - 3)^2$
(c) $a + a + a + \dots + a = Na$
(d) $f_1 X_1 + f_2 X_2 + f_3 X_3 + f_4 X_4 + f_5 X_5$
(e) $(X_1 - a) + (X_2 - a) + (X_3 - a) = X_1 + X_2 + X_3 - 3a$

3.2 用求和符号表达下列各式:

$$\begin{aligned} \text{(a)} & X_1^2 + X_2^2 + X_3^2 + \dots + X_{10}^2; \\ \text{(b)} & (X_1 + Y_1) + (X_2 + Y_2) + \dots + (X_8 + Y_8); \\ \text{(c)} & f_1 X_1^3 + f_2 X_2^3 + \dots + f_{20} X_{20}^3; \\ \text{(d)} & a_1 b_1 + a_2 b_2 + a_3 b_3 + \dots + a_N b_N; \\ \text{(e)} & f_1 X_1 Y_1 + f_2 X_2 Y_2 + f_3 X_3 Y_3 + f_4 X_4 Y_4. \end{aligned}$$

解  (a) $\sum_{j=1}^{10} X_j^2$; (c) $\sum_{j=1}^{20} f_j X_j^3$; (e) $\sum_{j=1}^4 f_j X_j Y_j$.
(b) $\sum_{j=1}^8 (X_j + Y_j)$; (d) $\sum_{j=1}^N a_j b_j$;

- 3.3 证明 $\sum_{j=1}^N (aX_j + bY_j - cZ_j) = a \sum_{j=1}^N X_j + b \sum_{j=1}^N Y_j - c \sum_{j=1}^N Z_j$, 其中 a, b 和 c 是任意常数.

证明

$$\begin{aligned} & \sum_{j=1}^N (aX_j + bY_j - cZ_j) \\ &= (aX_1 + bY_1 - cZ_1) + (aX_2 + bY_2 - cZ_2) + \cdots + (aX_N + bY_N - cZ_N) \\ &= (aX_1 + aX_2 + \cdots + aX_N) + (bY_1 + bY_2 + \cdots + bY_N) - (cZ_1 + cZ_2 + \cdots + cZ_N) \\ &= a(X_1 + X_2 + \cdots + X_N) + b(Y_1 + Y_2 + \cdots + Y_N) - c(Z_1 + Z_2 + \cdots + Z_N) \\ &= a \sum_{j=1}^N X_j + b \sum_{j=1}^N Y_j - c \sum_{j=1}^N Z_j \end{aligned}$$

或简略记为

$$\sum (aX + bY - cZ) = a \sum X + b \sum Y - c \sum Z.$$

- 3.4 两个变量 X 和 Y , 假设 $X_1=2, X_2=-5, X_3=4, X_4=-8, Y_1=-3, Y_2=-8, Y_3=10, Y_4=6$. 计算 (a) $\sum X$, (b) $\sum Y$, (c) $\sum XY$, (d) $\sum X^2$, (e) $\sum Y^2$, (f) $(\sum X)(\sum Y)$, (g) $\sum XY^2$, (h) $\sum (X+Y)(X-Y)$.

解 注意每式中, X 和 Y 的下标 j 都被省略了, \sum 应理解为 $\sum_{j=1}^4$. 例如, $\sum_{j=1}^4 X_j$ 简写为 $\sum X$.

$$(a) \sum X = 2 + (-5) + 4 + (-8) = 2 - 5 + 4 - 8 = -7$$

$$(b) \sum Y = (-3) + (-8) + 10 + 6 = -3 - 8 + 10 + 6 = 5$$

$$\begin{aligned} (c) \sum XY &= 2 \times (-3) + (-5) \times (-8) + 4 \times 10 + (-8) \times 6 \\ &= -6 + 40 + 40 - 48 = 26 \end{aligned}$$

$$(d) \sum X^2 = 2^2 + (-5)^2 + 4^2 + (-8)^2 = 4 + 25 + 16 + 64 = 109$$

$$(e) \sum Y^2 = (-3)^2 + (-8)^2 + 10^2 + 6^2 = 9 + 64 + 100 + 36 = 209$$

$$(f) \text{ 根据 (a) 和 (b), } (\sum X)(\sum Y) = (-7) \times 5 = -35. \text{ 注意, } (\sum X)(\sum Y) \neq \sum XY$$

$$(g) \sum XY^2 = 2(-3)^2 + (-5)(-8)^2 + 4 \times 10^2 + (-8) \times 6^2 = -190$$

$$(h) \sum (X+Y)(X-Y) = \sum (X^2 - Y^2) = \sum X^2 - \sum Y^2 = 109 - 209 = -100$$

- 3.5 如果 $\sum_{j=1}^6 X_j = -4, \sum_{j=1}^6 X_j^2 = 10$, 计算: (a) $\sum_{j=1}^6 (2X_j + 3)$, (b) $\sum_{j=1}^6 X_j(X_j - 1)$, (c) $\sum_{j=1}^6 (X_j - 5)^2$.

$$\text{解 (a) } \sum_{j=1}^6 (2X_j + 3) = \sum_{j=1}^6 2X_j + \sum_{j=1}^6 3 = 2 \sum_{j=1}^6 X_j + 6 \times 3 = 2(-4) + 18 = 10$$

$$(b) \sum_{j=1}^6 X_j(X_j - 1) = \sum_{j=1}^6 (X_j^2 - X_j) = \sum_{j=1}^6 X_j^2 - \sum_{j=1}^6 X_j = 10 - (-4) = 14$$

$$\begin{aligned} (c) \sum_{j=1}^6 (X_j - 5)^2 &= \sum_{j=1}^6 (X_j^2 - 10X_j + 25) = \sum_{j=1}^6 X_j^2 - 10 \sum_{j=1}^6 X_j + 25 \times 6 \\ &= 10 - 10(-4) + 25 \times 6 = 200 \end{aligned}$$

如果需要, 我们可在不产生误解的情况下省略下标 j , 用 \sum 代替 $\sum_{j=1}^6$.

算术平均

- 3.6 一个学生 6 门功课的成绩分别为 84, 91, 72, 68, 87 和 78, 求 6 门成绩的算术平均.

$$\text{解 } \bar{X} = \frac{\sum X}{N} = \frac{84 + 91 + 72 + 68 + 87 + 78}{6} = \frac{480}{6} = 80$$

有时, 平均与算术平均是同义的. 但是严格地说, 这并不正确, 因为除了算术平均外, 还有其他定义的平均数.

- 3.7 一个科学家记录了一个圆柱体直径的 10 个测量值: 3.88, 4.09, 3.92, 3.97, 4.02, 3.95, 4.03, 3.92, 3.98 和 4.06 厘米(cm). 求这些测量值的算术平均.

解 $\bar{X} = \frac{\sum X}{N}$

$$= \frac{3.88 + 4.09 + 3.92 + 3.97 + 4.02 + 3.95 + 4.03 + 3.92 + 3.98 + 4.06}{10}$$

$$= \frac{39.82}{10} = 3.98 \text{ cm}$$

- 3.8 如下 Minitab 产生的数据显示了 30 个 Internet 用户每星期在线所花的时间以及 30 个时间的平均值. 你能判断这个平均值具有代表性吗? Minitab 软件操作如下:

MTB>print cl

Data Display

time

3	4	4	5	5	5	5	5	5	6
6	6	6	7	7	7	7	7	8	8
9	10	10	10	10	10	10	12	55	60

MTB>mean cl

Column Mean

Mean of time = 10.400

解 平均值 10.4 小时不具代表性. 在这 30 个时间数字里有 21 个是一位数字, 但平均值为 10.4 小时. 这个平均值最大的缺点在于它会在很大程度上受到极值的影响.

- 3.9 求数字 5, 3, 6, 5, 4, 5, 2, 8, 6, 5, 4, 8, 3, 4, 5, 4, 8, 2, 5 和 4 的算术平均.

解 解法一

$$\bar{X} = \frac{\sum X}{N}$$

$$= \frac{5 + 3 + 6 + 5 + 4 + 5 + 2 + 8 + 6 + 5 + 4 + 8 + 3 + 4 + 5 + 4 + 8 + 2 + 5 + 4}{20}$$

$$= \frac{96}{20} = 4.8$$

解法二 这里有 6 个 5, 2 个 3, 2 个 6, 5 个 4, 2 个 2 和 3 个 8, 因此

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N} = \frac{6 \times 5 + 2 \times 3 + 2 \times 6 + 5 \times 4 + 2 \times 2 + 3 \times 8}{6 + 2 + 2 + 5 + 2 + 3}$$

$$= \frac{96}{20} = 4.8$$

- 3.10 在 100 个数中, 有 20 个 4, 40 个 5, 30 个 6, 其他为 7, 求这些数的算术平均.

解 $\bar{X} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N} = \frac{20 \times 4 + 40 \times 5 + 30 \times 6 + 10 \times 7}{100} = \frac{530}{100} = 5.30$

- 3.11 一个学生数学、物理、英语和卫生学的期末成绩分别为 82, 86, 90 和 70. 如果这些课程各自的学分分别为 3, 5, 3 和 1, 求一个恰当的平均分.

解 用加权算术平均, 权取为学分数.

$$\bar{X} = \frac{\sum wX}{\sum w} = \frac{3 \times 82 + 5 \times 86 + 3 \times 90 + 1 \times 70}{3 + 5 + 3 + 1} = 85$$

- 3.12 一公司有 80 个员工, 其中 60 个人每小时挣 10 美元, 20 个人每小时挣 13 美元.

(a) 求每小时平均所得.

(b) 如果 60 个员工的平均小时工资为 10.00 美元, 答案会与(a)一样吗?

(c) 你认为平均小时工资具代表性吗?

解 (a) $\bar{X} = \frac{\sum fX}{N} = \frac{60 \times 10.00 + 20 \times 13.00}{60 + 20} = 10.75$ 美元

(b) 结果会一样. 下面给出证明. 假设 f_1 个数的平均值为 m_1 , f_2 个数的平均值为 m_2 . 我们要证明所有数的平均值为

$$\bar{X} = \frac{f_1 m_1 + f_2 m_2}{f_1 + f_2}$$

假设 f_1 个数的和为 M_1 , f_2 个数的和为 M_2 . 根据算术平均的定义,

$$m_1 = \frac{M_1}{f_1}, \quad m_2 = \frac{M_2}{f_2}$$

或 $M_1 = f_1 m_1$, $M_2 = f_2 m_2$. 由于 $(f_1 + f_2)$ 个数的和为 $(M_1 + M_2)$, 这些数的算术平均为

$$\bar{X} = \frac{M_1 + M_2}{f_1 + f_2} = \frac{f_1 m_1 + f_2 m_2}{f_1 + f_2}$$

这正是我们要证明的. 结果很容易推广.

(c) 由于大多数员工小时工资为 10.00 美元, 这与 10.75 美元相差不多, 因此我们认为 10.75 美元是有“代表性”的小时工资. 必须记住, 当我们把数值数据相加为一个数时 (如在计算平均值时那样), 我们可能会造成一些误差. 然而, 结果不会像习题 3.8 那样产生误解.

实际中, 为了安全起见, 应该给出关于数据平均值的“差距”或“变差”. 这称为数据的**离差**. 离差的不同度量将在第四章介绍.

- 3.13** 四组学生分别由 15, 20, 10 和 18 人组成, 每一组的平均体重分别为 162, 148, 153 和 140 磅. 求所有学生的平均体重.

解 $\bar{X} = \frac{\sum fX}{\sum f} = \frac{15 \times 162 + 20 \times 148 + 10 \times 153 + 18 \times 140}{15 + 20 + 10 + 18} = 150$ 磅

- 3.14** 如果农业和非农业收入者的平均年薪分别为 25 000 美元和 35 000 美元, 那么两组人合起来的平均年薪是否为 30 000 美元?

解 只有当农业和非农业收入者人数相等时, 两组人合起来的平均年薪是 30 000 美元. 要知道真正的平均年薪, 就必须知道每一组的相对人数. 假设有 10% 的人是农业收入者. 此时, 平均值应为 $0.10 \times 25\,000 + 0.90 \times 35\,000 = 34\,000$ 美元. 如果两组人数相等, 平均值应为 $0.50 \times 25\,000 + 0.50 \times 35\,000 = 30\,000$ 美元.

- 3.15** 用表 2.1 身高的频数分布来求 XYZ 大学 100 个男同学的平均身高.

解 数据已经归纳在表 3.1 中. 注意, 这里身高 60~62 英寸, 63~65 英寸等等都被认为身高是 61 英寸, 64 英寸等等. 问题转化为求 100 个学生的身高, 其中 5 个学生身高 61 英寸, 18 个学生身高 64 英寸等等.

特别是在数据很多并且组数很多的情况下, 计算会很复杂. 恰当的短技术会减少计算量, 见习题 3.20 和 3.22.

表 3.1

身高(英寸)	组中点(X)	频数(f)	fX
60~62	61	5	305
63~65	64	18	1152
66~68	67	42	2814
69~71	70	27	1890
72~74	73	8	584
		$N = \sum f = 100$	$\sum fX = 6745$

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N} = \frac{6745}{100} = 67.45 \text{ 英寸}$$

算术平均的性质

3.16 证明 X_1, X_2, \dots, X_N 与它们的平均值 \bar{X} 的差的和等于零.

证明 设 $d_1 = X_1 - \bar{X}, d_2 = X_2 - \bar{X}, \dots, d_N = X_N - \bar{X}$.

$$\begin{aligned}\text{差的和} &= \sum d_j = \sum (X_j - \bar{X}) = \sum X_j - N\bar{X} \\ &= \sum X_j - N\left(\frac{\sum X_j}{N}\right) = \sum X_j - \sum X_j = 0\end{aligned}$$

其中, $\sum_{j=1}^N$ 由 \sum 代替. 如需要, 可在不产生误解的情况下, 省略 X_j 中的下标 j .

3.17 如果 $Z_1 = X_1 + Y_1, Z_2 = X_2 + Y_2, \dots, Z_N = X_N + Y_N$, 证明 $\bar{Z} = \bar{X} + \bar{Y}$.

证明 根据定义,

$$\bar{X} = \frac{\sum X}{N} \quad \bar{Y} = \frac{\sum Y}{N} \quad \bar{Z} = \frac{\sum Z}{N}$$

因此

$$\bar{Z} = \frac{\sum Z}{N} = \frac{\sum (X + Y)}{N} = \frac{\sum X + \sum Y}{N} = \frac{\sum X}{N} + \frac{\sum Y}{N} = \bar{X} + \bar{Y}$$

其中, $\sum_{j=1}^N$ 由 \sum 代替, X, Y 和 Z 的下标 j 省略了.

3.18 (a) 如果 N 个数 X_1, X_2, \dots, X_N 与任意数 A 的差为 $d_1 = X_1 - A, d_2 = X_2 - A, \dots, d_N = X_N - A$, 证明:

$$\bar{X} = A + \frac{\sum_{j=1}^N d_j}{N} = A + \frac{\sum d}{N}$$

(b) X_1, X_2, \dots, X_K 的频数分别为 f_1, f_2, \dots, f_K , 并且 $d_1 = X_1 - A, d_2 = X_2 - A, \dots, d_K = X_K - A$, 证明: 此时(a)中的结论变为

$$\bar{X} = A + \frac{\sum_{j=1}^K f_j d_j}{\sum_{j=1}^K f_j} = A + \frac{\sum f d}{N}$$

其中

$$\sum_{j=1}^K f_j = \sum f = N$$

证明 (a) 证法一 由于 $d_j = X_j - A, X_j = A + d_j$, 我们得到

$$\bar{X} = \frac{\sum X_j}{N} = \frac{\sum (A + d_j)}{N} = \frac{\sum A + \sum d_j}{N} = \frac{NA + \sum d_j}{N} = A + \frac{\sum d_j}{N}$$

其中, $\sum_{j=1}^N$ 由 \sum 代替.

证法二 省略 d 和 X 的下标, 我们有 $d = X - A$, 或 $X = A + d$. 因此根据习题3.17, 且由于一组常数 A 的平均数仍然是 A , 我们得到

$$\bar{X} = \bar{A} + \bar{d} = A + \frac{\sum d}{N}$$

(b)

$$\begin{aligned}\bar{X} &= \frac{\sum_{j=1}^K f_j X_j}{\sum_{j=1}^K f_j} = \frac{\sum f_j X_j}{N} = \frac{\sum f_j (A + d_j)}{N} = \frac{\sum A f_j + \sum f_j d_j}{N} \\ &= \frac{A \sum f_j + \sum f_j d_j}{N} = \frac{AN + \sum f_j d_j}{N} = A + \frac{\sum f_j d_j}{N} = A + \frac{\sum f d}{N}\end{aligned}$$

注意在形式上结果可以从(a)得到, 只要把 d_j 换成 $f_j d_j$, 求和从 $j=1$ 到 K 换成 $j=1$ 到 N . 此结果等价于 $\bar{X} = A + \bar{d}$, 其中 $\bar{d} = (\sum fd)/N$.

从分类资料中计算算术平均

3.19 用习题 3.18(a)的方法求 5, 8, 11, 9, 12, 6, 14 和 10 的算术平均, 其中 A 取:(a)9;(b)20.

解 (a) 给定数字与 9 的差分别为 -4, -1, 2, 0, 3, -3, 5 和 1, 这些差的和为 $\sum d = -4 - 1 + 2 + 0 + 3 - 3 + 5 + 1 = 3$. 因此

$$\bar{X} = A + \frac{\sum d}{N} = 9 + \frac{3}{8} = 9.375$$

(b) 给定数字与 20 的差分别为 -15, -12, -9, -11, -8, -14, -6 和 -10, $\sum d = -85$. 因此

$$\bar{X} = A + \frac{\sum d}{N} = 20 + \frac{-85}{8} = 9.375$$

3.20 用习题 3.18(b)的方法求 XYZ 大学 100 个男同学的平均身高(见习题 3.15).

解 数据已经整理在表 3.2 中. 尽管任何组中值都可作为 A , 我们取 A 为组中值 67(它的频数最大). 这里的计算比习题 3.15 中的简单一些. 要使得计算更简便, 我们就要采用习题 3.22 中的方法, 那里的差(表 3.2 的第二列)都是组距大小的整倍数.

表 3.2

组中值(X)	差 $d = X - A$	频数(f)	fd
61	-6	5	-30
64	-3	18	-54
$A \rightarrow 67$	0	42	0
70	3	27	81
73	6	8	48
		$N = \sum f = 100$	$\sum fd = 45$

$$\bar{X} = A + \frac{\sum fd}{N} = 67 + \frac{45}{100} = 67.45 \text{ 英寸}$$

3.21 $d_j = X_j - A$ 表示在一频数分布中任一组中值 X_j 与一给定组中值 A 的差. 证明如果所有的组距有相同的大小 c , 那么(a)这些差都是 c 的倍数(即, $d_j = cu_j$, $u_j = 0, \pm 1, \pm 2, \dots$);(b)平均值的计算公式可写为

$$\bar{X} = A + \left(\frac{\sum fu}{N} \right) c$$

证明 (a) 习题 3.20 的表 3.2 已经说明了结论, 其中第二列中的数都是组距大小 $c = 3$ 英寸的整数倍.

为了证明在一般情况下结论也成立, 设 X_1, X_2, X_3, \dots 是连续组中值, 它们共同的差为 c , 因此, $X_2 = X_1 + c, X_3 = X_2 + c$, 一般地, $X_j = X_1 + (j-1)c$. 任意两个组中值 X_p 和 X_q 的差为

$$X_p - X_q = [X_1 + (p-1)c] - [X_1 + (q-1)c] = (p-q)c$$

显然这是 c 的倍数.

(b) 根据(a), 所有组中值与任一给定组中值的差是 c 的倍数(即, $d_j = cu_j$). 利用习题 3.18(b), 得到

$$\bar{X} = A + \frac{\sum fd_j}{N} = A + \frac{\sum f_j(cu_j)}{N} = A + c \frac{\sum f_j u_j}{N} = A + \left(\frac{\sum fu}{N} \right) c$$

注意此式等价于 $\bar{X} = A + c\bar{u}$, 只要在 $\bar{X} = A + \bar{d}$ 中用 $c\bar{u}$ 代替 \bar{d} , 即 $c\bar{u}$ 代替 \bar{d} (见习题 3.18).

3.22 用习题 3.21(b)的结论求 XYZ 大学 100 个男同学的平均身高(见习题 3.20).

解 数据已经整理在表 3.3 中. 这种方法称为编码法, 可尽可能地使用.

表 3.3

X	u	f	fu
61	-2	5	-10
64	-1	18	-18
$A \rightarrow 67$	0	42	0
70	1	27	27
73	2	8	16
		$N = 100$	$\sum fu = 15$

$$\bar{X} = A + \left(\frac{\sum fu}{N} \right)_c = 67 + \frac{15}{100} \times 3 = 67.45 \text{ 英寸}$$

3.23 根据表 2.5 的频数分布用(a)长方法;(b)编码法计算 P&R 公司 65 个员工的平均周薪.

解 表 3.4 和 3.5 分别给出了(a)和(b)的解.

表 3.4

$X(\text{美元})$	f	fX
255.00	8	2040.00
265.00	10	2650.00
275.00	16	4400.00
285.00	14	3990.00
295.00	10	2950.00
305.00	5	1525.00
315.00	2	630.00
$N = 65$		$\sum fX = 18,185.00$

表 3.5

$X(\text{美元})$	u	f	fu
255.00	-2	8	-16
265.00	-1	10	-10
$A \rightarrow 275.00$	0	16	0
285.00	1	14	14
295.00	2	10	20
305.00	3	5	15
315.00	4	2	8
		$N = 65$	$\sum fu = 31$

由于实际上组中值是 254.995 美元, 264.995 美元等, 而非 255.00 美元, 265.00 美元等, 因此在这些表中有必要引进误差. 如果在表 3.4 中用真实的组中值进行计算, 那么 \bar{X} 不是 279.77 美元, 而是 279.76 美元, 两者的差可忽略.

$$\bar{X} = \frac{\sum fX}{N} = \frac{18,185.00}{65} = 279.77 \text{ 美元}$$

$$\bar{X} = A + \left(\frac{\sum fu}{N} \right)_c = 275.00 + \frac{31}{65} \times 10 = 279.77 \text{ 美元}$$

3.24 根据表 2.9(d), 计算 P&R 公司 70 个员工的平均周薪.

解 这里, 组距大小不相同, 我们必须用长方法, 如表 3.6 所示.

表 3.6

$X(\text{美元})$	f	$fX(\text{美元})$
255.00	8	2040.00
265.00	10	2650.00
275.00	16	4400.00
285.00	15	4275.00
295.00	10	2950.00
310.00	8	2480.00
350.00	3	1050.00
$N = 70$		$\sum fX = 19,845.00$

$$\bar{X} = \frac{\sum fX}{N} = \frac{19\,845.00}{70} = 283.50 \text{ 美元}$$

中位数

- 3.25 如下 Minitab 产生的数据显示了 30 个 Internet 用户每周在线所花的时间以及 30 个时间的中位数. 验证这个中位数. 这个中位数具有代表性吗? 把你的结论与习题 3.8 的结论作个比较. Minitab 软件操作如下:

MTB>print cl

Data Display

time

3	4	4	5	5	5	5	5	5	6
6	6	6	7	7	7	7	7	8	8
9	10	10	10	10	10	10	12	55	60

MTB>median cl

Column Median

Median of time = 7.0000

解 这里两个中间值都为 7, 并且两个中间值的平均是 7. 习题 3.8 中, 平均值为 10.4 小时. 这里的中位数比平均值更具代表性.

- 3.26 在一个大城市里, 有 15 个地方记录 ATM 每天的交易量. 这些数据为: 35, 49, 225, 50, 30, 65, 40, 55, 52, 76, 48, 325, 47, 32 和 60. 求 (a) 交易量的中位数, (b) 交易量的平均值.

解 (a) 按顺序这些数排列为: 30, 32, 35, 40, 47, 48, 49, 50, 52, 55, 60, 65, 76, 225 和 325. 由于这里有奇数个数, 因此只有一个中间值 50, 即为所求的中位数.

(b) 15 个值的和为 1189. 平均值为 $1189/15 = 79.267$.

注意, 中位数不受两个极值 225 和 325 的影响, 而平均值会受到它们的影响. 在这个例子中, 中位数较好地描述了 ATM 每天交易量的平均值.

- 3.27 如果有 (a) 85; (b) 150 个数排列在一个数组里, 如何找出它们的中位数?

解 (a) 由于 85 个数是奇数个数, 只有一个中间值, 在它前面有 42 个数, 在它后面有 42 个数. 因此中位数是数组中第 43 个数.

(b) 由于 150 个数是偶数个数, 有两个中间值, 在它们前面有 74 个数, 在它们后面有 74 个数. 这两个中间值是数组里第 75 和第 76 个数, 它们的算术平均就是所求的中位数.

- 3.28 根据习题 2.8(a) 表 2.7 的频数分布 (重新设计为表 3.7); (b) 原始数据分别求州立大学 40 个男同学体重的中位数.

解 (a) **解法一 (插值法)** 假设表 3.7 的体重频数分布是连续的. 在此例中, 中位数是有总频数 $(40/2 = 20)$ 一半的体重低于它, 一半的体重高于它的体重值.

前三组频数的和为 $3 + 5 + 9 = 17$. 为了得到 20, 我们至少在第四组的 12 个值中取 3 个. 由于第四组 145~153 对应体重为 144.5 到 153.5, 因此中位数为

$$144.5 + \frac{3}{12} \times (153.5 - 144.5) = 144.5 + \frac{3}{12} \times 9 = 146.8 \text{ 磅}$$

表 3.7

体重 (磅)	频数
118~126	3
127~135	5
136~144	9
145~153	12
154~162	5
163~171	4
172~180	2
总计	40

解法二 (公式法) 由于前三组和前四组的频数和分别为 $3 + 5 + 9 = 17$ 和 $3 + 5 + 9 + 12 = 29$, 显然中位数在第四组, 因此第四组就是中位数组.

$$L_1 = \text{中位数组的下组界} = 144.5$$

$$N = \text{数据的数目} = 40$$

$$\left(\sum f \right)_1 = \text{中位数组前各组的频数和} = 3 + 5 + 9 = 17$$

$$f_m = \text{中位数组组频数} = 12$$

$$c = \text{中位数组组距宽度} = 9$$

因此,

$$\text{中位数} = L_1 + \left[\frac{\frac{N}{2} - (\sum f)_1}{f_m} \right] c = 144.5 + \frac{40/2 - 17}{12} \times 9 = 146.8 \text{ 磅}$$

(b)原始数据排列成数组:

119, 125, 126, 128, 132, 135, 135, 135, 136, 138, 138, 140, 140, 142, 142, 144, 144, 145, 145, 146, 146, 147, 147, 148, 149, 150, 150, 152, 153, 154, 156, 157, 158, 161, 163, 164, 165, 168, 173, 176

中位数等于数组中第 20 和第 21 个体重值的算术平均, 等于 146 磅.

3.29 从(a)直方图, (b)百分率卵形图中求习题 3.28 体重的中位数.

解 (a)图 3-3(a)是习题 3.28 体重对应的直方图. 直线 LM 把直方图分为面积相等的两部分, 中位数就是直线 LM 对应的横坐标. 由于在直方图中面积与频数是对应的, 直线 LM 左边和右边的频数为总频数的一半, 即 20. 因此 $AMLD$ 和 $MBEL$ 的面积分别对应于频数 3 和 9. $AM = \frac{3}{12} AB = \frac{3}{12} \times 9 = 2.25$, 中位数的值为 $144.5 + 2.25 = 146.75$ 或舍入至最近的十分位数 146.8 磅. 这个值也可近似地从图中读出.

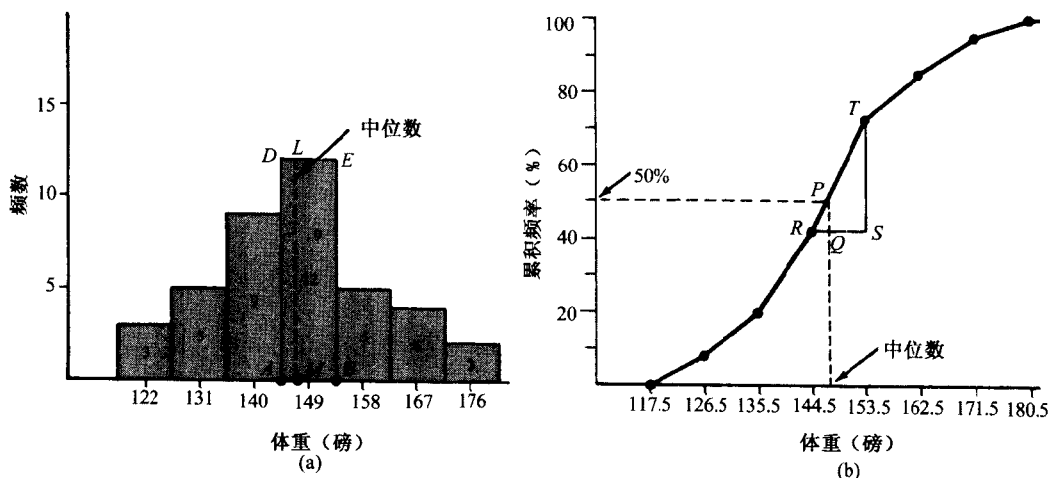


图 3-3

(b)图 3-3(b)是习题 3.28 体重对应的累积频率多边形(或百分率卵形线). 中位数是曲线上 P 点的横坐标, P 点的纵坐标是 50%. 为了计算结果, 看相似三角形 PQR 和 TSR ,

$$\frac{RQ}{RS} = \frac{PQ}{ST} \quad \text{或} \quad \frac{RQ}{9} = \frac{50\% - 42.5\%}{72.5\% - 42.5\%} = \frac{1}{4} \quad \text{得} \quad RQ = \frac{9}{4} = 2.25$$

因此, 中位数 $= 144.5 + RQ = 144.5 + 2.25 = 146.75$ 磅

或舍入至最近的十分位数 146.8 磅. 这个值也可近似地从图中读出.

3.30 求 P&R 公司 65 个员工工资的中位数(见习题 2.3).

解 这里 $N = 65$, $N/2 = 32.5$. 由于前两组和前三组的频数和分别为 $8 + 10 = 18$ 和 $8 + 10 + 16 = 34$, 中位数组是第三组. 应用公式,

$$\text{中位数} = L_1 + \left[\frac{\frac{N}{2} - (\sum f)_1}{f_m} \right] c = 269.995 + \frac{32.5 - 18}{16} \times 10.00 = 279.06 \text{ 美元}$$

众数

3.31 求(a)3, 5, 2, 6, 5, 9, 5, 2, 8, 6, (b)51.6, 48.7, 50.3, 49.5, 48.9 的平均数, 中位数和众数.

解 (a)排列成数组: 2, 2, 3, 5, 5, 5, 6, 6, 8 和 9.

$$\text{平均值} = \frac{1}{10} (2 + 2 + 3 + 5 + 5 + 5 + 6 + 6 + 8 + 9) = 5.1$$

$$\text{中位数} = \text{两个中间数的算术平均} = \frac{1}{2}(5+5) = 5$$

众数 = 发生次数最多的数 = 5

(b) 排列成数组: 48.7, 48.9, 49.5, 50.3, 51.6.

$$\text{平均值} = \frac{1}{5}(48.7 + 48.9 + 49.5 + 50.3 + 51.6) = 49.8$$

中位数 = 中间数 = 49.5

众数 = 发生次数最多的数(这里不存在)

3.32 写出根据频数分布中的数据求众数的公式.

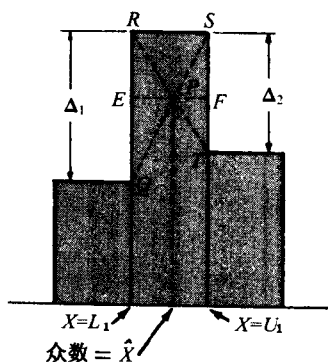


图 3-4

解 假设图 3-4 表示频数分布直方图的三个矩形, 中间的矩形与众数组对应, 并且组距有相等的大小.

我们定义众数是 QS 和 RT 连线交点 P 的横坐标 \bar{X} .

$X = L_1$ 和 $X = U_1$ 分别代表众数组的下组界和上组界, Δ_1 和 Δ_2 分别代表众数组频数与左右两组频数的差.

根据相似三角形 PQR 和 PST, 我们有

$$\frac{EP}{RQ} = \frac{PF}{ST} \quad \text{或} \quad \frac{\bar{X} - L_1}{\Delta_1} = \frac{U_1 - \bar{X}}{\Delta_2}$$

那么

$$\Delta_2(\bar{X} - L_1) = \Delta_1(U_1 - \bar{X}), \quad \Delta_2\bar{X} - \Delta_2L_1 = \Delta_1U_1 - \Delta_1\bar{X}$$

$$(\Delta_1 + \Delta_2)\bar{X} = \Delta_1U_1 + \Delta_2L_1$$

或

$$\bar{X} = \frac{\Delta_1U_1 + \Delta_2L_1}{\Delta_1 + \Delta_2}$$

由于 $U_1 = L_1 + c$, 其中 c 是组距大小, 上式变为

$$\bar{X} = \frac{\Delta_1(L_1 + c) + \Delta_2L_1}{\Delta_1 + \Delta_2} = \frac{(\Delta_1 + \Delta_2)L_1 + \Delta_1c}{\Delta_1 + \Delta_2} = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)c$$

这个结果有一个有趣的解释. 如果一条抛物线经过图 3-4 三个矩形上端的中点, 那么这条抛物线顶点的横坐标与上述得到的众数一致.

3.33 用习题 3.32 得到的公式求 P&R 公司 65 个员工工资的众数(见习题 3.23).

解 这里 $L_1 = 269.995$ 美元, $\Delta_1 = 16 - 10 = 6$, $\Delta_2 = 16 - 14 = 2$, $c = 10.00$ 美元, 因此,

$$\text{众数} = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)c = 269.995 + \frac{6}{2+6} \times 10.00 = 277.50 \text{ 美元}$$

均值, 中位数和众数之间的经验关系

3.34 (a) 用经验公式: 均值 - 众数 = 3(均值 - 中位数), 求 P&R 公司 65 个员工工资的众数.

(b) 用这里的结果与习题 3.33 的结果作比较.

解 (a) 根据习题 3.23 和 3.30 我们知道均值 = 279.77 美元, 中位数 = 279.06 美元. 因此,

$$\text{众数} = \text{均值} - 3(\text{均值} - \text{中位数}) = 279.77 - 3(279.77 - 279.06) = 277.64 \text{ 美元}.$$

(b) 习题 3.33 中的工资众数是 277.50 美元, 与经验结果有较好的一致性.

几何平均

3.35 求 3, 5, 6, 6, 7, 10 和 12 的 (a) 几何平均; (b) 算术平均.

解 (a) 几何平均 = $G = \sqrt[7]{3 \times 5 \times 6 \times 6 \times 7 \times 10 \times 12} = \sqrt[7]{453600}$. 用常用对数, $\log G = \frac{1}{7} \log 453600 = \frac{1}{7} \times 5.6567 = 0.8081$, $G = 6.43$ (舍入至最近的百分位). 也可用计算器计算.

$$\begin{aligned} \text{另解} \quad \log G &= \frac{1}{7} (\log 3 + \log 5 + \log 6 + \log 6 + \log 7 + \log 10 + \log 12) \\ &= \frac{1}{7} (0.4771 + 0.6990 + 0.7782 + 0.7782 + 0.8451 + 1.0000 + 1.0792) \end{aligned}$$

$$= 0.8081$$

$$G = 6.43$$

(b) 算术平均 $\bar{X} = \frac{1}{7}(3+5+6+6+7+10+12) = 7$. 这说明了一组不等正数的几何平均值小于算术平均值.

3.36 数 X_1, X_2, \dots, X_K 发生的频数分别为 f_1, f_2, \dots, f_K , 其中 $f_1 + f_2 + \dots + f_K = N$ 为总频数.

(a) 求几何平均 G ;

(b) 写出 $\log G$ 的表达式;

(c) 如何用得到的结果求频数分布中分类资料的几何平均?

解 (a) $G = \sqrt[N]{\underbrace{X_1 X_1 \cdots X_1}_{f_1 \text{ 次}} \underbrace{X_2 X_2 \cdots X_2}_{f_2 \text{ 次}} \cdots \underbrace{X_K X_K \cdots X_K}_{f_K \text{ 次}}} = \sqrt[N]{X_1^{f_1} X_2^{f_2} \cdots X_K^{f_K}}$

其中, $N = \sum f$. 这常称为**加权几何平均**.

$$\begin{aligned} \text{(b)} \log G &= \frac{1}{N} \log(X_1^{f_1} X_2^{f_2} \cdots X_K^{f_K}) = \frac{1}{N} (f_1 \log X_1 + f_2 \log X_2 + \cdots + f_K \log X_K) \\ &= \frac{1}{N} \sum_{j=1}^K f_j \log X_j = \frac{\sum f \log X}{N} \end{aligned}$$

这里我们假设所有的数都是正的; 否则对数没有定义.

注意, 一组正数的几何平均的对数等于这组数对数的算术平均.

(c) 如果假设 X_1, X_2, \dots, X_K 是组中值, f_1, f_2, \dots, f_K 是相应的频数, 那么上述结果可用来求分类数据的几何平均值.

3.37 一年中每夸特牛奶对每条面包的价格比率为 3.00, 而在下一年中这个比率为 2.00.

(a) 求两年间牛奶对面包价格比率的算术平均;

(b) 求两年间面包对牛奶价格比率的算术平均;

(c) 讨论用算术平均来计算平均比率的可行性;

(d) 讨论用几何平均来计算平均比率的合理性.

解 (a) 牛奶对面包价格比率的算术平均 $= \frac{1}{2} \times (3.00 + 2.00) = 2.50$.

(b) 由于第一年牛奶对面包价格比率是 3.00, 因此面包对牛奶价格比率是 $1/3.00 = 0.333$. 同理, 第二年面包对牛奶价格比率是 $1/2.00 = 0.500$, 所以, 面包对牛奶价格比率的算术平均 $= \frac{1}{2} (0.333 + 0.500) = 0.417$.

(c) 如果算术平均是恰当的平均数, 那么牛奶对面包价格比率的算术平均是面包对牛奶价格比率的算术平均的倒数. 然而, $1/0.417 = 2.40 \neq 2.50$. 因此, 这里用算术平均来计算平均比率是不可行的.

(d) 牛奶对面包价格比率的几何平均 $= \sqrt{3.00 \times 2.00} = \sqrt{6.00}$.

面包对牛奶价格比率的几何平均 $= \sqrt{0.333 \times 0.500} = \sqrt{0.0167} = 1/\sqrt{6.00}$.

由于这两个平均值互为倒数, 因此对这一类型的问题我们认为更适合用几何平均来计算平均比率.

3.38 在某种环境下一种细菌数量在三天内由 1000 增加到 4000. 求每天的平均增长率.

解 由于从 1000 增加到 4000 增长了 300%, 有人可能认为每天的平均增长率是 $300\% / 3 = 100\%$. 然而, 这意味着在第一天细菌数从 1000 增加到 2000, 第二天细菌数从 2000 增加到 4000, 第三天细菌数从 4000 增加到 8000, 显然与事实不符.

为了求出正确的结论, 我们把这个增长率记为 r .

$$\text{一天后细菌数} = 1000 + 1000r = 1000(1+r)$$

$$\text{两天后细菌数} = 1000(1+r) + 1000(1+r)r = 1000(1+r)^2$$

$$\text{三天后细菌数} = 1000(1+r)^2 + 1000(1+r)^2 r = 1000(1+r)^3$$

最后一个表达式应等于 4000. 因此 $1000(1+r)^3 = 4000$, $(1+r)^3 = 4$, $1+r = \sqrt[3]{4}$, $r = \sqrt[3]{4} - 1 = 1.587 - 1 = 0.587$, $r = 58.7\%$.

一般地, 如果开始量是 P , 单位时间增长率是一常数 r , 在 n 个单位时间后, 数量变为

$$A = P(1+r)^n$$

这称为复利公式(见习题 3.94 和 3.95).

调和平均

3.39 求 3, 5, 6, 6, 7, 10 和 12 的调和平均 H .

$$\begin{aligned}\text{解 } \frac{1}{H} &= \frac{1}{N} \sum \frac{1}{X} = \frac{1}{7} \left(\frac{1}{3} + \frac{1}{5} + \frac{1}{6} + \frac{1}{6} + \frac{1}{7} + \frac{1}{10} + \frac{1}{12} \right) \\ &= \frac{1}{7} \left(\frac{140 + 84 + 70 + 70 + 60 + 42 + 35}{420} \right) = \frac{501}{2940} \\ H &= \frac{2940}{501} = 5.87\end{aligned}$$

通常把分数先表示为小数形式会简便一些, 如

$$\begin{aligned}\frac{1}{H} &= \frac{1}{7} (0.3333 + 0.2000 + 0.1667 + 0.1667 + 0.1429 + 0.1000 + 0.0833) \\ &= \frac{1.1929}{7}\end{aligned}$$

$$\text{此时, } H = \frac{7}{1.1929} = 5.87.$$

与习题 3.35 作比较后发现, 一些不全等的正数的调和平均小于它们的几何平均, 因此也就小于算术平均.

3.40 在连续的 4 年里, 一个家庭主妇分别以每加仑(gal)0.80 美元, 0.90 美元, 1.05 美元和 1.25 美元的价格购买火炉燃油. 求 4 年间油的平均价格.

解 情况一 假设这个家庭主妇每年购买相同数量的燃油, 比如 1000 加仑. 那么

$$\text{平均价格} = \frac{\text{总价}}{\text{总量}} = \frac{800 + 900 + 1050 + 1250}{4000} = 1.00 \text{ 美元 / 加仑}$$

这与每加仑的价格的算术平均是一样的, 即

$$\frac{1}{4} (0.80 + 0.90 + 1.05 + 1.25) = 1.00 \text{ 美元 / 加仑}$$

如果每年购买 x 加仑燃油, 这个结果依然成立.

情况二 假设这个家庭主妇每年支出相等的钱购买燃油, 比如 1000 美元, 那么

$$\text{平均价格} = \frac{\text{总价}}{\text{总量}} = \frac{4000}{1250 + 1111 + 952 + 800} = 0.975 \text{ 美元 / 加仑}$$

这与每加仑的价格的调和平均是一样的, 即

$$\frac{4}{\frac{1}{0.80} + \frac{1}{0.90} + \frac{1}{1.05} + \frac{1}{1.25}} = 0.975$$

如果每年 y 美元购买燃油, 这个结果依然成立.

尽管两种平均值是在不同情况下计算得到的, 但它们的平均过程都是正确的.

注意, 如果每一年购买的燃油数量不等, 那么情况一中的普通算术平均就应用加权算术平均来代替. 同理, 如果每一年支付不等的钱购买燃油, 那么情况二中的普通调和平均就应用加权调和平均.

3.41 一辆车以时速 25 英里行驶 25 英里, 以时速 50 英里行驶 25 英里, 以时速 75 英里行驶 25 英里. 求三个速度的算术平均和调和平均. 哪一个正确呢?

解 平均速度等于总路程除以总时间,

$$\frac{75}{1 + \frac{1}{2} + \frac{1}{3}} = 40.9 \text{ 英里 / 小时}$$

3 个速度的算术平均值为

$$\frac{25 + 50 + 75}{3} = 50 \text{ 英里 / 小时}$$

调和平均如下求解:

$$\frac{1}{H} = \frac{1}{N} \sum \frac{1}{X} = \frac{1}{3} \left(\frac{1}{25} + \frac{1}{50} + \frac{1}{75} \right) = \frac{11}{450} \quad H = \frac{450}{11} = 40.9 \text{ 英里 / 小时}$$

调和平均是平均速度正确的度量.

均方根

3.42 求 3, 5, 6, 6, 7, 10 和 12 的均方根.

解 均方根 = $\sqrt{\frac{3^2 + 5^2 + 6^2 + 6^2 + 7^2 + 10^2 + 12^2}{7}} = \sqrt{57} = 7.55$

3.43 证明两个不等正数 a, b 的均方根大于它们的几何平均.

证明 我们要证明 $\sqrt{\frac{1}{2}(a^2 + b^2)} > \sqrt{ab}$. 假如此不等式成立, 两边同时平方得:
 $\frac{1}{2}(a^2 + b^2) > ab$, 即, $a^2 + b^2 > 2ab$, $a^2 - 2ab + b^2 > 0$, 或 $(a - b)^2 > 0$. 由于任何不为零的实数的平方是正的, 因此最后的不等式是正确的.

上述步骤倒推可得到本题的证明. 因为 $(a - b)^2 > 0$, 所以 $a^2 + b^2 > 2ab$, $\frac{1}{2}(a^2 + b^2) > ab$, 最后得到 $\sqrt{\frac{1}{2}(a^2 + b^2)} > \sqrt{ab}$.

注意, $\sqrt{\frac{1}{2}(a^2 + b^2)} = \sqrt{ab}$ 当且仅当 $a = b$.

四分位数, 十分位数和百分位数

3.44 求 P&R 公司 65 个员工工资的 (a) 四分位数 Q_1, Q_2, Q_3 ; (b) 十分位数 D_1, D_2, \dots, D_9 (见习题 2.3).

解 (a) Q_1 是从第一组数起第 $N/4 = 65/4 = 16.25$ 个工资数. 由于第一组有 8 个数据, 必须从第二组中取 $8.25 = 16.25 - 8$ 个数据. 用线性插值法, 得到 $Q_1 = 259.995 + \frac{8.25}{10} \times 10.00 = 268.25$ 美元
 Q_2 是从第一组数起第 $2N/4 = N/2 = 65/2 = 32.5$ 个工资数. 由于前两组有 18 个数据, 必须从第三组中取 $32.5 - 18 = 14.5$ 个数据, 因此

$$Q_2 = 269.995 + \frac{14.5}{16} \times 10.00 = 279.06 \text{ 美元}$$

Q_3 是从第一组数起第 $3N/4 = \frac{3}{4} \times 65 = 48.75$ 个工资数. 由于前四组有 48 个数据, 必须从第三组中取 $48.75 - 48 = 0.75$ 个数据, 因此

$$Q_3 = 289.995 + \frac{0.75}{10} \times 10.00 = 290.75 \text{ 美元}$$

因此 25% 的员工工资不多于 268.25 美元, 50% 的员工工资不多于 279.06 美元, 75% 的员工工资不多于 290.75 美元.

(b) 第一, 第二, \dots , 第九个十分位数分别是第一组数起的 $N/10, 2N/10, \dots, 9N/10$ 个数. 因此

$$D_1 = 249.995 + \frac{6.5}{8} \times 10.00 = 258.12 \text{ 美元}$$

$$D_2 = 259.995 + \frac{5}{10} \times 10.00 = 265.00 \text{ 美元}$$

$$D_3 = 269.995 + \frac{1.5}{16} \times 10.00 = 270.94 \text{ 美元}$$

$$D_4 = 269.995 + \frac{8}{16} \times 10.00 = 275.00 \text{ 美元}$$

$$D_5 = 269.995 + \frac{14.5}{16} \times 10.00 = 279.06 \text{ 美元}$$

$$D_6 = 279.995 + \frac{5}{14} \times 10.00 = 283.57 \text{ 美元}$$

$$D_7 = 279.995 + \frac{11.5}{14} \times 10.00 = 288.21 \text{ 美元}$$

$$D_8 = 289.995 + \frac{4}{10} \times 10.00 = 294.00 \text{ 美元}$$

$$D_9 = 299.995 + \frac{0.5}{5} \times 10.00 = 301.00 \text{ 美元}$$

所以 10% 的员工工资不多于 258.12 美元, 20% 的员工工资不多于 265.00 美元, …… , 90% 的员工工资不多于 301.00 美元.

这里第五个十分位数是中位数. 第二、第四、第六和第八个十分位数把分布分为 5 个相等的部分, 称为**五分位数**, 经常应用在实践中.

3.45 求习题 3.44 分布的 (a) 第 35 个和 (b) 第 60 个百分位数.

解 (a) 第三十五个百分位数记为 P_{35} , 是从第一组数起第 $35N/100 = 35 \times 65/100 = 22.75$ 个数据. 如习题 3.44,

$$P_{35} = 269.995 + \frac{4.75}{16} \times 10.00 = 272.97 \text{ 美元}$$

这意味着 35% 的员工工资不多于 272.97 美元.

(b) 第六十个百分位数 $P_{60} = 279.995 + \frac{5}{14} \times 10.00 = 283.57$ 美元. 这与第六个十分位数是相等的.

3.46 证明习题 3.44 和 3.45 的结论可从百分率卵形线中得到.

证明

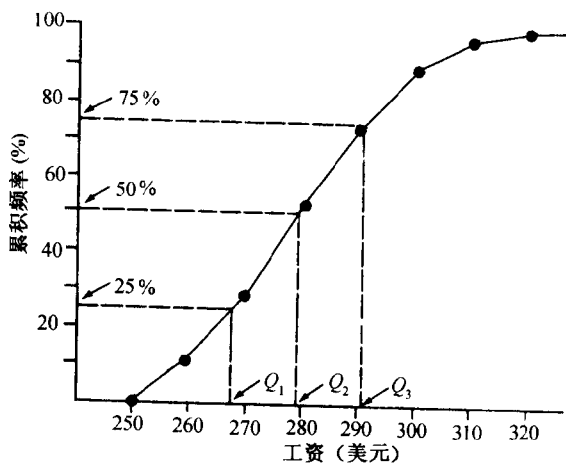


图 3-5

根据习题 3.44 和 3.45 的数据得到的相应百分率卵形线如图 3-5.

第一个四分位数是卵形线上纵坐标为 25% 的点的横坐标. 同样的, 第二和第三个四分位数分别是卵形线上纵坐标为 50% 和 75% 的点的横坐标.

十分位数和百分位数都可用类似的方法得到. 例如, 第七个十分位数和第三十五个百分位数分别是卵形线上纵坐标为 70% 和 35% 的点的横坐标.

补充习题

求和记号

3.47 写出下列各式的展开式:

$$\begin{aligned} \text{(a)} \sum_{j=1}^4 (X_j + 2); & \quad \text{(c)} \sum_{j=1}^3 U_j (U_j + 6); & \quad \text{(e)} \sum_{j=1}^4 4X_j Y_j. \\ \text{(b)} \sum_{j=1}^5 f_j X_j^2; & \quad \text{(d)} \sum_{k=1}^N (Y_k^2 - 4); \end{aligned}$$

3.48 用求和记号表达下列各式:

$$\begin{aligned} \text{(a)} & (X_1 + 3)^2 + (X_2 + 3)^2 + (X_3 + 3)^2; \\ \text{(b)} & f_1(Y_1 - a)^2 + f_2(Y_2 - a)^2 + \cdots + f_{15}(Y_{15} - a)^2; \\ \text{(c)} & (2X_1 - 3Y_1) + (2X_2 - 3Y_2) + \cdots + (2X_N - 3Y_N); \\ \text{(d)} & (X_1/Y_1 - 1)^2 + (X_2/Y_2 - 1)^2 + \cdots + (X_8/Y_8 - 1)^2; \end{aligned}$$

$$(e) \frac{f_1 a_1^2 + f_2 a_2^2 + \cdots + f_{12} a_{12}^2}{f_1 + f_2 + \cdots + f_{12}}.$$

$$3.49 \quad \text{证明} \quad \sum_{j=1}^N (X_j - 1)^2 = \sum_{j=1}^N X_j^2 - 2 \sum_{j=1}^N X_j + N.$$

$$3.50 \quad \text{证明} \quad \sum (X + a)(Y + b) = \sum XY + a \sum Y + b \sum X + Nab, \text{ 其中 } a, b \text{ 是常数. 求和符号表示什么?}$$

3.51 两个变量 U 和 V , 假设 $U_1 = 3, U_2 = -2, U_3 = 5, V_1 = -4, V_2 = -1, V_3 = 6$. 求

$$(a) \sum UV, (b) \sum (U + 3)(V - 4), (c) \sum V^2,$$

$$(d) (\sum U)(\sum V)^2, (e) \sum UV^2, (f) \sum (U^2 - 2V^2 + 2), (g) \sum (U/V).$$

$$3.52 \quad \text{已知} \quad \sum_{j=1}^4 X_j = 7, \sum_{j=1}^4 Y_j = -3, \sum_{j=1}^4 X_j Y_j = 5, \text{ 求 } (a) \sum_{j=1}^4 (2X_j + 5Y_j); (b) \sum_{j=1}^4 (X_j - 3)(2Y_j + 1).$$

算术平均

3.53 一个学生五门功课的成绩分别为 85, 76, 93, 82 和 96. 求成绩的算术平均.

3.54 一个心理学家测得一个人对某些刺激的反应时间分别为 0.53, 0.46, 0.50, 0.49, 0.52, 0.53, 0.44 和 0.55 秒. 求这些反应时间的算术平均.

3.55 一组数由 6 个 6, 7 个 7, 8 个 8, 9 个 9 和 10 个 10 组成. 求这些数的算术平均.

3.56 一个学生物理课的实验、报告和背诵部分的成绩分别为 71, 78 和 89.

(a) 如果这些成绩的权分别为 2, 4 和 5. 求一个恰当的平均分.

(b) 如果权相等, 那么平均分是多少?

3.57 3 个经济学教师所教班级平均考试成绩分别为 79, 74 和 82, 班级人数分别为 32, 25 和 17 人. 求三个班级的平均分.

3.58 一个公司支付给所有员工的平均年薪是 36 000 美元. 支付给男员工和女员工的平均年薪分别为 34000 美元和 40000 美元. 求男女员工人数的百分比.

3.59 表 3.8 显示的是某公司生产的缆绳所能承载的最大负荷量, 单位为短吨 (1 短吨 = 2000 磅). 用 (a) “长方法”; (b) 编码法求平均最大负荷量.

表 3.8

最大负荷量(短吨)	缆绳数
9.3~9.7	2
9.8~10.2	5
10.3~10.7	12
10.8~11.2	17
11.3~11.7	14
11.8~12.2	6
12.3~12.7	3
12.8~13.2	1
总计	60

3.60 用 (a) “长方法”; (b) 编码法求表 3.9 中数据的 \bar{X} .

表 3.9

X	462	480	498	516	534	552	570	588	606	624
f	98	75	56	42	30	21	15	11	6	2

3.61 表 3.10 显示了一公司生产的铆钉顶端直径的分布. 求平均直径.

3.62 计算表 3.11 中数据的算术平均.

表 3.10

直径(cm)	频数
0.7247~0.7249	2
0.7250~0.7252	6
0.7253~0.7255	8
0.7256~0.7258	15
0.7259~0.7261	42
0.7262~0.7264	68
0.7265~0.7267	49
0.7268~0.7270	25
0.7271~0.7273	18
0.7274~0.7276	12
0.7277~0.7279	4
0.7280~0.7282	1
总计	250

表 3.11

组	频数
10~15 以下	3
15~20 以下	7
20~25 以下	16
25~30 以下	12
30~35 以下	9
35~40 以下	5
40~45 以下	2
总计	54

3.63 计算习题 2.20 中 400 个初中生每周看电视的平均时间。

3.64 (a)用习题 2.27 得到的频数分布计算滚珠的平均直径。

(b)用原始数据直接计算平均值并把它与(a)中的结论作比较,解释差异。

中位数

3.65 求(a)5,4,8,3,7,2,9;(b)18.3,20.6,19.3,22.4,20.2,18.8,19.7,20.0 的算术平均和中位数。

3.66 求习题 3.53 中成绩的中位数。

3.67 求习题 3.54 中反应时间的中位数。

3.68 求习题 3.55 中数据的中位数。

3.69 根据习题 3.59 中表 3.8 计算缆绳最大负荷量的中位数。

3.70 根据习题 3.60 表 3.9 计算分布的中位数 \tilde{x} 。

3.71 根据习题 3.61 中表 3.10 计算铆钉顶端直径的中位数。

3.72 根据习题 3.62 中表 3.11 计算分布的中位数。

3.73 表 3.12 给出了 1993 年因为心脏疾病而死亡的人数。求死亡年龄的中位数。

表 3.12

年龄组	死亡数(千人)
全部	743.3
低于 1	0.7
1~4	0.3
5~14	0.3
15~24	1.0
25~34	3.5
35~44	13.1
45~54	32.7
55~64	72.0
65~74	158.1
75~84	234.0
高于 85	227.6

来源:美国国家健康统计中心,美国生死统计,年鉴。

3.74 根据习题 2.31 的数据求美国人口年龄的中位数。

3.75 计算习题 2.20 中 400 个初中生每周看电视时间的中位数。

众数

3.76 求(a)7,4,10,9,15,12,7,9,7;(b)8,11,4,3,2,5,10,6,4,1,10,8,12,6,5,7 的算术平均,中位数和众

数.

- 3.77 求习题 3.53 成绩的众数.
- 3.78 求习题 3.54 反应时间的众数.
- 3.79 求习题 3.55 中数据的众数.
- 3.80 求习题 3.59 中缆绳的最大负荷量的众数.
- 3.81 根据习题 3.60 表 3.9 求分布的众数 \bar{X} .
- 3.82 根据习题 3.61 表 3.10 求铆钉顶端直径的众数.
- 3.83 求习题 3.62 中分布的众数.
- 3.84 求习题 2.20 中 400 个初中生每周看电视时间的众数.
- 3.85 (a)表 2.15 年龄组的众数是什么?
(b)表 3.12 年龄组的众数是什么?
- 3.86 用本章公式(9)和(10)计算下列习题中分布的众数.比较你用公式得到的答案.
(a)习题 3.59;(b)习题 3.61;(c)习题 3.62;(d)习题 2.20.
- 3.87 证明习题 3.32 结尾所给的结论.

几何平均

- 3.88 求(a)4.2 和 16.8;(b)3.00 和 6.00 的几何平均.
- 3.89 求 2,4,8,16,32 的几何平均 G 和算术平均 \bar{X} .
- 3.90 求(a)3,5,8,3,7,2;(b)28.5,73.6,47.2,31.5,64.8 的几何平均.
- 3.91 求(a)习题 3.59;(b)习题 3.60 中分布的几何平均.证明这些数据的几何平均小于等于算术平均.
- 3.92 如果一种商品的价格在 4 年间翻了一番,求年平均增长率.
- 3.93 1980 和 1996 年美国的人口分别为 226.5×10^6 和 266.0×10^6 ,用习题 3.38 所给公式,回答下列问题.
(a)年平均增长率是多少?
(b)估计 1985 年的人口数.
(c)如果从 1996 到 2000 年人口的年平均增长率与(a)中的相同,2000 年的人口数将是多少?
- 3.94 1000 元本金以年利率 8% 进行投资.如果最初的本金不被取出,那么 6 年后本息和是多少?
- 3.95 如果在习题 3.94 中利率按季以复利计息(即每 3 个月增长 2%),6 年后本息和是多少?
- 3.96 若两个数的算术平均是 9.0,几何平均是 7.2.求这两个数.

调和平均

- 3.97 求(a)2,3 和 6,(b)3.2,5.2,4.8,6.1 和 4.2 的调和平均.
- 3.98 求 0,2,4 和 6 的(a)算术平均,(b)几何平均,(c)调和平均.
- 3.99 如果 X_1, X_2, X_3, \dots 代表频数分布的组中值,相应的组频数分别为 f_1, f_2, f_3, \dots ,证明分布的调和平均 H 可表示为

$$\frac{1}{H} = \frac{1}{N} \left(\frac{f_1}{X_1} + \frac{f_2}{X_2} + \frac{f_3}{X_3} + \dots \right) = \frac{1}{N} \sum \frac{f}{X}$$

其中, $N = f_1 + f_2 + \dots = \sum f$.

- 3.100 根据习题 3.99 求(a)习题 3.59;(b)习题 3.60 中分布的调和平均,并与习题 3.91 作比较.
- 3.101 城市 A,B 和 C 相互间距离相等.一个乘车者以 30 英里/小时的速度从 A 到 B,以 40 英里/小时的速度从 B 到 C,以 50 英里/小时的速度从 C 到 A.求整个行程的平均速度.
- 3.102 (a)一架飞机分别以 v_1, v_2 和 v_3 英里/小时的速度飞行 d_1, d_2 和 d_3 英里.证明平均速度 V 满足下式:

$$\frac{d_1 + d_2 + d_3}{V} = \frac{d_1}{v_1} + \frac{d_2}{v_2} + \frac{d_3}{v_3}$$

这是一个加权调和平均.

- (b)如果 $d_1 = 2500, d_2 = 1200, d_3 = 500, v_1 = 500, v_2 = 400, v_3 = 250$.求 V .
- 3.103 证明两个正数 a 和 b 的几何平均(a)小于等于算术平均;(b)大于等于调和平均.多于两个数的情况呢?

均方根

- 3.104 求(a)11, 23 和 35;(b)2.7, 3.8, 3.2 和 4.3 的均方根.
- 3.105 证明两个正数 a 和 b 的均方根(a)大于等于算术平均;(b)大于等于调和平均. 多于两个数的情况呢?
- 3.106 写出从分类数据中求得均方根的公式, 并把它应用到一个已考察过的频数分布中.

四分位数, 十分位数和百分位数

- 3.107 表 3.13 给出了高等数学期末考试成绩的频数分布.(a)求分布的四分位数;(b)解释每个分位数的含义.

表 3.13

分数	学生数
90~100	9
80~89	32
70~79	43
60~69	21
50~59	11
40~49	3
30~39	1
总计	120

- 3.108 求(a)习题 3.59;(b)习题 3.60 中分布的四分位数 Q_1, Q_2, Q_3 . 解释每个分位数的含义.
- 3.109 用六种不同的统计术语描述钟形频数曲线的平衡点或中心值.
- 3.110 根据习题 3.59 中的数据求(a) P_{10} ;(b) P_{90} ;(c) P_{25} ;(d) P_{75} , 并解释它们的含义.
- 3.111 (a)所有的四分位数和十分位数可以表示为百分位数吗? 请解释原因.
(b)所有的四分位数都可以表示为十分位数吗? 请解释原因.
- 3.112 根据习题 3.107 的数据求(a)组内成绩居于前 25% 中的最低成绩;(b)组内成绩居于后 20% 中的最高成绩. 根据百分位数解释你的结论.
- 3.113 用(a)百分率直方图;(b)百分率频数多边形;(c)百分率卵形线解释习题 3.107 中的结论.
- 3.114 用习题 3.108 的结论解答习题 3.113.
- 3.115 (a)写出一个类似于本章(8)式那样的公式对频数分布计算百分位数.
(b)用这个公式来求习题 3.110 的解并说明公式的用途.

第四章 标准差和其他表示离差的度量

离差或变差

数值数据围绕其平均值分布的分数与集中程度称为数据的**离差或变差**. 根据不同度量, 可以定义不同的离差(或变差), 最常用的有全距, 平均偏差, 半内四分位数间距, 10~90 百分位数间距和标准差.

全距

一组数的**全距**是这组数中最大的数与最小的数的差.

例 1 一组数 2, 3, 3, 5, 5, 5, 8, 10, 12 的全距为 $12 - 2 = 10$. 有时, 全距也可简单地用最大与最小的数来表示. 例如, 在此例中, 全距可表示为 2 到 12, 或 $2 \sim 12$.

平均偏差

N 个数 X_1, X_2, \dots, X_N 的**平均偏差**简记为 **MD**, 并定义如下:

$$\text{平均偏差(MD)} = \frac{\sum_{j=1}^N |X_j - \bar{X}|}{N} = \frac{\sum |X - \bar{X}|}{N} = \overline{|X - \bar{X}|} \quad (1)$$

其中 \bar{X} 是这 N 个数的算术平均(自本章起, 大多将算术平均简称为均值), $|X_j - \bar{X}|$ 是 X_j 与 \bar{X} 差的绝对值. (一个数的**绝对值**是这个数去掉相应的符号后的数, 用两条竖直线放在数字两边来表示. 因此, $|-4| = 4$, $|+3| = 3$, $|6| = 6$, $|-0.84| = 0.84$.)

例 2 求 2, 3, 6, 8, 11 的平均偏差.

$$\begin{aligned}\bar{X} &= \frac{2+3+6+8+11}{5} = 6 \\ MD &= \frac{|2-6| + |3-6| + |6-6| + |8-6| + |11-6|}{5} \\ &= \frac{|-4| + |-3| + |0| + |2| + |5|}{5} \\ &= \frac{4+3+0+2+5}{5} = 2.8\end{aligned}$$

如果 X_1, X_2, \dots, X_k 分别发生 f_1, f_2, \dots, f_k 次, 平均偏差可写为

$$MD = \frac{\sum_{j=1}^K f_j |X_j - \bar{X}|}{N} = \frac{\sum f |X - \bar{X}|}{N} = \overline{|X - \bar{X}|} \quad (2)$$

其中 $N = \sum_{j=1}^K f_j = \sum f$. 这个形式对分类数据很有用, 此时 X_j 是第 j 组的组中值, f_j 是对应的组频数.

有时, 可以根据数据与中位数或其他平均值之差的绝对值来定义平均偏差. 求和形式

$\sum_{j=1}^N |X_j - a|$ 有一个有趣的性质: 当 a 是中位数时和最小(即关于中位数的平均偏差是最小值).

注意, 使用术语**平均绝对偏差**比**平均偏差**要更恰当些.

半内四分位数间距

一组数据的**半内四分位数间距**或**半内四分距**用 Q 表示, 定义为

$$Q = \frac{Q_3 - Q_1}{2} \quad (3)$$

其中 Q_1 和 Q_3 分别是数据的第一和第三个四分位数(见习题 4.6 和 4.7). 有时也会使用四分位数间距 $Q_3 - Q_1$, 但是半内四分位数间距更多地作为离差的度量.

10~90 百分位数间距

一组数的 10~90 百分位数间距定义为

$$10 \sim 90 \text{ 百分位数间距} = P_{90} - P_{10} \quad (4)$$

其中 P_{10} 和 P_{90} 是数据的第 10 个和第 90 个百分位数(见习题 4.8). 半 10~90 百分位数间距 $\frac{1}{2}(P_{90} - P_{10})$ 不经常使用.

标准差

N 个数 X_1, X_2, \dots, X_N 的标准差用 s 表示, 定义为

$$s = \sqrt{\frac{\sum_{j=1}^N (X_j - \bar{X})^2}{N}} = \sqrt{\frac{\sum (X - \bar{X})^2}{N}} = \sqrt{\frac{\sum x^2}{N}} = \sqrt{(X - \bar{X})^2} \quad (5)$$

其中, x 代表 X_j 与均值 \bar{X} 的差. 因此 s 是 X_j 与 \bar{X} 偏差的均方根, 也常称为均方根差.

如果 X_1, X_2, \dots, X_K 分别发生 f_1, f_2, \dots, f_K 次, 标准差可写为

$$s = \sqrt{\frac{\sum_{j=1}^K f_j (X_j - \bar{X})^2}{N}} = \sqrt{\frac{\sum f (X - \bar{X})^2}{N}} = \sqrt{\frac{\sum f x^2}{N}} = \sqrt{(X - \bar{X})^2} \quad (6)$$

其中 $N = \sum_{j=1}^K f_j = \sum f$. 这个形式对分类数据很有用.

有时在定义一个样本数据的标准差时, 表达式(5)和(6)的分母常用 $(N-1)$ 代替 N , 因为这样产生的值是总体标准差的较好估计. 当 N 较大时 ($N > 30$), 这两种定义没有什么区别. 同样, 我们可以根据第一个定义得到的标准差乘以 $\sqrt{N/(N-1)}$ 而得到这个较好的估计值. 因此我们必须掌握好(5)和(6)式.

方差

一组数据的方差定义为标准差的平方, 用(5)和(6)中的 s^2 表示.

当有必要区别总体的标准差和从这个总体抽取的样本的标准差时, 我们常用符号 s 代表后者, σ (Σ 的小写字母) 表示前者, 因此 s^2 和 σ^2 分别表示样本方差和总体方差.

计算标准差的快捷方法

(5)和(6)式可分别写为下列等价形式:

$$s = \sqrt{\frac{\sum_{j=1}^N X_j^2}{N} - \left[\frac{\sum_{j=1}^N X_j}{N} \right]^2} = \sqrt{\frac{\sum X^2}{N} - \left[\frac{\sum X}{N} \right]^2} = \sqrt{\overline{X^2} - \bar{X}^2} \quad (7)$$

$$s = \sqrt{\frac{\sum_{j=1}^K f_j X_j^2}{N} - \left[\frac{\sum_{j=1}^K f_j X_j}{N} \right]^2} = \sqrt{\frac{\sum f X^2}{N} - \left[\frac{\sum f X}{N} \right]^2} = \sqrt{\overline{X^2} - \bar{X}^2} \quad (8)$$

其中 $\overline{X^2}$ 表示 X 平方的均值, \bar{X}^2 表示 X 的均值的平方(见习题 4.12 和 4.14).

如果 $d_j = X_j - A$ 表示 X_j 与某个任意常数 A 的差, (7)和(8)式可分别表示为

$$s = \sqrt{\frac{\sum_{j=1}^N d_j^2}{N} - \left[\frac{\sum_{j=1}^N d_j}{N} \right]^2} = \sqrt{\frac{\sum d^2}{N} - \left[\frac{\sum d}{N} \right]^2} = \sqrt{d^2 - \bar{d}^2} \quad (9)$$

$$s = \sqrt{\frac{\sum_{j=1}^K f_j d_j^2}{N} - \left[\frac{\sum_{j=1}^K f_j d_j}{N} \right]^2} = \sqrt{\frac{\sum f d^2}{N} - \left[\frac{\sum f d}{N} \right]^2} = \sqrt{d^2 - \bar{d}^2} \quad (10)$$

(见习题 4.15 和 4.17).

当数据整理成组距大小相等 ($=c$) 的频数分布时, 记 $d_j = cu_j$ 或 $X_j = A + cu_j$, 则 (10) 式变为

$$s = c \sqrt{\frac{\sum_{j=1}^K f_j u_j^2}{N} - \left[\frac{\sum_{j=1}^K f_j u_j}{N} \right]^2} = c \sqrt{\frac{\sum f u^2}{N} - \left[\frac{\sum f u}{N} \right]^2} = c \sqrt{u^2 - \bar{u}^2} \quad (11)$$

最后的公式提供求标准差的快捷方法, 当分类数据组距大小相等时可以用来计算. 这称为**编码法**, 并且它与第三章计算分类数据的算术平均值的算法是很相似的 (见习题 4.16~4.19).

标准差的性质

1. 一般的标准差可定义为

$$s = \sqrt{\frac{\sum_{j=1}^N (X_j - a)^2}{N}}$$

其中 a 是除算术平均外还可以是其他的平均值. 由第三章的性质 2 可知, 所有这些标准差的最小值在 $a = \bar{X}$ 时取到, 这就是用 (5) 式定义标准差的原因. 这个性质的证明见习题 4.27.

2. 对于正态分布 (见第七章), 可以证明 (如图 4-1):

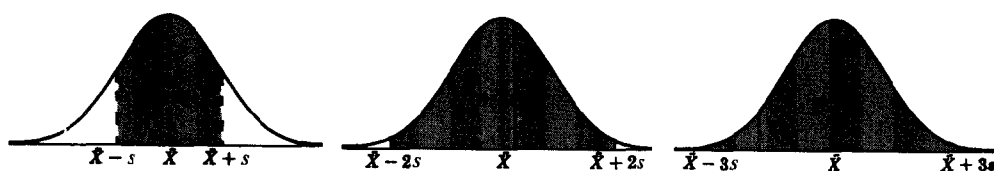


图 4-1

(a) 68.27% 的数据包含在 $\bar{X} - s$ 和 $\bar{X} + s$ 之间 (即均值两边一个标准差内).

(b) 95.45% 的数据包含在 $\bar{X} - 2s$ 和 $\bar{X} + 2s$ 之间 (即均值两边两个标准差内).

(c) 99.73% 的数据包含在 $\bar{X} - 3s$ 和 $\bar{X} + 3s$ 之间 (即均值两边三个标准差内).

对于微斜的分布, 上述百分比近似成立 (见习题 4.24).

3. 假设分别由 N_1 和 N_2 个数构成的两组数 (或者总频数分别为 N_1 和 N_2 的两个频数分布), 它们的方差分别为 s_1^2 和 s_2^2 , 有相同的均值 \bar{X} . 那么两组数 (或两个频数分布) 的**组合的或合并的方差**定义如下:

$$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2} \quad (1.2)$$

这是两个方差的加权算术平均. 这个结论可推广到三个或更多的数组.

Charlier 检验

用编码法计算均值和标准差的 Charlier 检验利用了下列性质:

$$\sum f(u+1) = \sum fu + \sum f = \sum fu + N$$

$$\begin{aligned}\sum f(u+1)^2 &= \sum f(u^2 + 2u + 1) = \sum fu^2 + 2\sum fu + \sum f \\ &= \sum fu^2 + 2\sum fu + N\end{aligned}$$

(见习题 4.20).

Sheppard 方差修正

在数据分组时,标准差的计算会出现误差(分组误差).为了调整分组误差,我们使用公式:

$$\text{修正方差} = \text{分组数据方差} - \frac{c^2}{12} \quad (13)$$

其中 c 是组距大小.修正量 $c^2/12$ 称为 **Sheppard 修正**.它通常用在两个方向的“尾部”趋于零的连续变量的分布中.

统计学家常考虑何时和是否使用 Sheppard 修正.因为它经常会**过分修正**,所以在应用前应进行细致的研究.本书中除非特别说明,我们将不会使用 Sheppard 修正.

离差度量间的经验关系

对于微斜的分布,我们有如下经验公式:

$$\text{平均偏差} = \frac{4}{5} \text{标准差}$$

$$\text{半内四分位数间距} = \frac{2}{3} \text{标准差}$$

这是因为对于正态分布我们发现平均偏差和半内四分位数间距分别是标准差的 0.7979 倍和 0.6745 倍.

绝对和相对离差,变异系数

从标准差或其他离差度量得到的真实变差或离差称为**绝对离差**.然而,在测量距离为 1000 英尺时,10 英寸的变差(或离差)产生的影响与测量距离是 20 英尺时同样的变差产生的影响是有很大区别的.这种影响的程度可用**相对离差**来减弱,它被定义为

$$\text{相对离差} = \frac{\text{绝对离差}}{\text{平均值}} \quad (14)$$

如果绝对离差是标准差 s ,平均值是均值 \bar{X} ,那么相对离差称为**变异系数**,用 V 表示,

$$\text{变异系数}(V) = \frac{s}{\bar{X}} \quad (15)$$

通常表示为百分数.其他可能的定义见习题 4.30.

注意,变异系数与所用的单位无关.因此,在不同单位的分布的比较中,这会是有益的.但当 \bar{X} 接近于零时,变异系数会失去效用.

标准化变量,标准分数

以变量 X 的标准差 s 为单位来度量 X 与其均值 \bar{X} 之间偏差的变量 z 称为**标准化变量**,即

$$z = \frac{X - \bar{X}}{s} \quad (16)$$

标准化变量是一个无量纲量(即与所用单位无关的量).

标准化变量的数值称为**标准分数**或 Z 分数.在比较分布时有很大用处(见习题 4.31).

习题与解答

全距

4.1 求(a)12, 6, 7, 3, 15, 10, 18, 5; (b)9, 3, 8, 8, 9, 8, 9, 18 的全距.

解 在两组数中, 均有全距 = 最大值 - 最小值 = $18 - 3 = 15$. 然而, 正如从数列(a)、(b)中看到的:

(a) 3, 5, 6, 7, 10, 12, 15, 18; (b) 3, 8, 8, 8, 9, 9, 9, 18

(a) 的离差比(b)的要大. 实际上, (b) 的大部分由 8 和 9 组成.

由于全距不能区别这两组数, 因此在此例中它不是离差的一个良好度量. 当极值存在时, 全距通常都不能很好地度量离差.

一个改进方法就是把极值 3 和 18 去掉. 此时(a)的全距是 $15 - 5 = 10$, (b)的全距是 $9 - 8 = 1$, 这就清楚的表明(a)的离差比(b)的大. 但这样一来就不符合全距的定义了. 半内四分位数间距和 10~90 百分位数间距就很好地弥补了这一问题的.

4.2 根据表 2.1 求 XYZ 大学学生身高的全距.

解 对于分类数据, 全距有两种定义方式.

解法一 全距 = 最高组组中值 - 最低组组中值
 $= 73 - 61 = 12$ 英寸

解法二 全距 = 最高组上组界 - 最低组下组界
 $= 74.5 - 59.5 = 15$ 英寸

第一种方法在某种程度上消除了极值的影响.

平均偏差

4.3 求习题 4.1 中数据的平均偏差.

解 (a) 均值为

$$\bar{X} = \frac{12 + 6 + 7 + 3 + 15 + 10 + 18 + 5}{8} = \frac{76}{8} = 9.5$$

平均偏差为

$$\begin{aligned} MD &= \frac{\sum |X - \bar{X}|}{N} \\ &= \frac{|12 - 9.5| + |6 - 9.5| + |7 - 9.5| + |3 - 9.5|}{8} \\ &\quad + \frac{|15 - 9.5| + |10 - 9.5| + |18 - 9.5| + |5 - 9.5|}{8} \\ &= \frac{2.5 + 3.5 + 2.5 + 6.5 + 5.5 + 0.5 + 8.5 + 4.5}{8} = \frac{34}{8} = 4.25 \end{aligned}$$

$$(b) \quad \bar{X} = \frac{9 + 3 + 8 + 8 + 9 + 8 + 9 + 18}{8} = \frac{72}{8} = 9$$

$$\begin{aligned} MD &= \frac{\sum |X - \bar{X}|}{N} \\ &= \frac{|9 - 9| + |3 - 9| + |8 - 9| + |8 - 9|}{8} \\ &\quad + \frac{|9 - 9| + |8 - 9| + |9 - 9| + |18 - 9|}{8} \\ &= \frac{0 + 6 + 1 + 1 + 0 + 1 + 0 + 9}{8} = 2.25 \end{aligned}$$

平均偏差显示了(b)的离差比(a)的小.

4.4 求 XYZ 大学 100 个男同学身高的平均偏差(见习题 3.20 表 3.2).

解 根据习题 3.20, $\bar{X} = 67.45$ 英寸. 把数据整理在表 4.1 中. 也可通过编码法来计算平均偏差(见习题 4.47).

$$MD = \frac{\sum f |X - \bar{X}|}{N} = \frac{226.50}{100} = 2.26 \text{ 英寸}$$

表 4.1

身高(英寸)	组中值(X)	$ X - \bar{X} = X - 67.45 $	频数(f)	$f X - \bar{X} $
60~62	61	6.45	5	32.25
63~65	64	3.45	18	62.10
66~68	67	0.45	42	18.90
69~71	70	2.55	27	68.85
72~74	73	5.55	8	44.40
			$N = \sum f = 100$	$\sum f X - \bar{X} = 226.50$

4.5 讨论习题 4.4 中学生身高落入区域(a) $\bar{X} \pm MD$, (b) $\bar{X} \pm 2MD$, (c) $\bar{X} \pm 3MD$ 中的百分比.

解 (a) $\bar{X} \pm MD = 67.45 \pm 2.26$ (即 65.19~69.71 英寸的区域). 这个区域包括第三组所有学生数 + $\frac{1}{3} \times (65.5 - 65.19)$ 倍第二组学生数 + $\frac{1}{3} \times (69.71 - 68.5)$ 倍第四组学生数 (由于组距大小是 3 英寸, 第二组上组界是 65.5 英寸, 第四组的下组界是 68.5 英寸). 在区域 $\bar{X} \pm MD$ 中的学生数为

$$42 + \frac{0.31}{3} \times 18 + \frac{1.21}{3} \times 27 = 42 + 1.86 + 10.89 = 54.75 \text{ 或 } 55$$

即占总数的 55%.

(b) $\bar{X} \pm 2MD = 67.45 \pm 2 \times 2.26 = 67.45 \pm 4.52$ (即 62.93~71.97 英寸的区域).

在区域 $\bar{X} \pm 2MD$ 中的学生数为

$$18 - \frac{62.93 - 62.5}{3} \times 18 + 42 + 27 + \frac{71.97 - 71.5}{3} \times 8 = 85.67 \text{ 或 } 86$$

即占总数的 86%.

(c) $\bar{X} \pm 3MD = 67.45 \pm 3 \times 2.26 = 67.45 \pm 6.78$ (即 60.67~74.23 英寸的区域).

在区域 $\bar{X} \pm 3MD$ 中的学生数为

$$5 - \frac{60.67 - 59.5}{3} \times 5 + 18 + 42 + 27 + \frac{74.23 - 74.5}{3} \times 8 = 97.33 \text{ 或 } 97$$

即占总数的 97%.

半内四分位数间距

4.6 求 XYZ 大学学生身高分布的半内四分位数间距 (见习题 4.4 表 4.1).

解 $Q_1 = 65.5 + \frac{2}{42} \times 3 = 65.64$ 英寸, $Q_3 = 68.5 + \frac{10}{27} \times 3 = 69.61$ 英寸, 则半内四分位数间距 (或半内四分距) 为 $Q = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2} \times (69.61 - 65.64) = 1.98$ 英寸. 注意, 这里有 50% 的数据落在 Q_1 和 Q_3 之间 (即, 50 个学生的身高介于 65.64 和 69.61 英寸之间).

我们把 $\frac{1}{2}(Q_3 + Q_1) = 67.63$ 英寸作为集中趋势的度量 (即平均身高). 可知有 50% 的身高落在区域 67.63 ± 1.98 英寸之间.

4.7 求 P&R 公司 65 个员工工资的半内四分位数间距 (见习题 2.3 表 2.5).

解 根据习题 3.44, $Q_1 = 268.25$ 元, $Q_3 = 290.75$ 元. 因此, 半内四分位数间距 $Q = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(290.75 - 268.25) = 11.25$ 元. 由于 $\frac{1}{2}(Q_3 + Q_1) = 279.50$ 元, 我们可下结论: 50% 员工工资落在区域 279.50 ± 11.25 元之间.

10~90 百分位数间距

4.8 求 XYZ 大学学生身高的 10~90 百分位数间距 (见表 2.1).

解 $P_{10} = 62.5 + \frac{5}{18} \times 3 = 63.33$ 英寸, $P_{90} = 68.5 + \frac{25}{27} \times 3 = 71.27$ 英寸. 因此 10~90 百分位数间距 $P_{90} - P_{10} = 71.27 - 63.33 = 7.94$ 英寸. 由于 $\frac{1}{2}(P_{90} + P_{10}) = 67.30$ 英寸, $\frac{1}{2}(P_{90} - P_{10}) = 3.97$

英寸, 我们知道有 80% 的学生身高落在 67.30 ± 3.97 英寸之间.

标准差

4.9 求习题 4.1 中每组数据的标准差 s .

解 (a) $\bar{X} = \frac{\sum X}{N} = \frac{12 + 6 + 7 + 3 + 15 + 10 + 18 + 5}{8} = \frac{76}{8} = 9.5$

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

$$= \left\{ \frac{1}{8} \left[(12 - 9.5)^2 + (6 - 9.5)^2 + (7 - 9.5)^2 + (3 - 9.5)^2 \right. \right. \\ \left. \left. + (15 - 9.5)^2 + (10 - 9.5)^2 + (18 - 9.5)^2 + (5 - 9.5)^2 \right] \right\}^{\frac{1}{2}}$$

$$= \sqrt{23.75} = 4.87$$

(b) $\bar{X} = \frac{9 + 3 + 8 + 8 + 9 + 8 + 9 + 18}{8} = \frac{72}{8} = 9$

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

$$= \left\{ \frac{1}{8} \left[(9 - 9)^2 + (3 - 9)^2 + (8 - 9)^2 + (8 - 9)^2 + (9 - 9)^2 \right. \right. \\ \left. \left. + (8 - 9)^2 + (9 - 9)^2 + (18 - 9)^2 \right] \right\}^{\frac{1}{2}}$$

$$= \sqrt{15} = 3.87$$

把以上结果与习题 4.3 的结果作比较, 我们可以看出标准差确实说明 (b) 的离差比 (a) 的小. 然而, 极值对标准差的影响比对平均偏差的影响大得多. 当然, 由于在计算标准差时离差进行平方运算, 因此这也是意料之中的.

4.10 由 Minitab 给出的习题 4.1 中两组数据的标准差在下面给出. 请将这里的结果与习题 4.9 的结果作比较.

MTB>print c1

set 1

12 6 7 3 15 10 18 5

MTB>print c2

set 2

9 3 8 8 9 8 9 18

MTB>standard deviation c1

Column Standard Deviation

Standard deviation of set1 = 5.21

MTB> standard deviation c2

Column Standard Deviation

Standard deviation of set2 = 4.14

解 Minitab 软件使用的公式为

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}}$$

因此习题 4.10 的标准差与习题 4.9 的不同. 如果我们在习题 4.9 的结果上乘以 $\sqrt{N/(N-1)}$, 那么就可以得到习题 4.10 的结果. $N = 8$, $\sqrt{N/(N-1)} = 1.069045$, 对于第一组数 $1.069045 \times 4.87 = 5.21$, 与 Minitab 给出的标准差相同. 同理, $1.069045 \times 3.87 = 4.14$, 与 Minitab 给出的标准差相同.

4.11 求 XYZ 大学 100 个男同学身高的标准差(见表 2.1).

解 根据习题 3.15, 3.20 或 3.22, $\bar{X} = 67.45$ 英寸. 数据整理在表 4.2 中.

表 4.2

身高 (英寸)	组中值 (X)	$X - \bar{X} = X - 67.45$	$(X - \bar{X})^2$	频数(f)	$f(X - \bar{X})^2$
60~62	61	-6.45	41.6025	5	208.0125
63~65	64	-3.45	11.9025	18	214.2450
66~68	67	-0.45	0.2025	42	8.5050
69~71	70	2.55	6.5025	27	175.5675
72~74	73	5.55	30.8025	8	246.4200
				$N = \sum f = 100$	$\sum f(X - \bar{X})^2 = 852.7500$

$$s = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}} = \sqrt{\frac{852.7500}{100}} = \sqrt{8.5275} = 2.92 \text{ 英寸}$$

根据分类资料计算标准差

4.12 (a)证明:

$$s = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2}$$

(b)用(a)中公式求 12, 6, 7, 3, 15, 10, 18, 5 的标准差.

解 (a)根据定义

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

因此

$$\begin{aligned} s^2 &= \frac{\sum (X - \bar{X})^2}{N} = \frac{\sum (X^2 - 2\bar{X}X + \bar{X}^2)}{N} = \frac{\sum X^2 - 2\bar{X} \sum X + N\bar{X}^2}{N} \\ &= \frac{\sum X^2}{N} - 2\bar{X} \frac{\sum X}{N} + \bar{X}^2 = \frac{\sum X^2}{N} - 2\bar{X}^2 + \bar{X}^2 = \frac{\sum X^2}{N} - \bar{X}^2 \\ &= \overline{X^2} - \bar{X}^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2 \end{aligned}$$

或

$$s = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2}$$

注意在上述求和中, 我们用 X 代替 X_j , \sum 代替 $\sum_{j=1}^N$.

另解

$$\begin{aligned} s^2 &= \overline{(X - \bar{X})^2} = \overline{X^2 - 2X\bar{X} + \bar{X}^2} = \overline{X^2} - \overline{2X\bar{X}} + \overline{\bar{X}^2} \\ &= \overline{X^2} - 2\bar{X} \bar{X} + \bar{X}^2 = \overline{X^2} - \bar{X}^2 \end{aligned}$$

$$(b) \overline{X^2} = \frac{\sum X^2}{N} = \frac{12^2 + 6^2 + 7^2 + 3^2 + 15^2 + 10^2 + 18^2 + 5^2}{8} = \frac{912}{8} = 114$$

$$\bar{X} = \frac{\sum X}{N} = \frac{12 + 6 + 7 + 3 + 15 + 10 + 18 + 5}{8} = \frac{76}{8} = 9.5$$

因此

$$s = \sqrt{\overline{X^2} - \bar{X}^2} = \sqrt{114 - 90.25} = \sqrt{23.75} = 4.87$$

此方法可与习题 4.9(a)作比较.

4.13 修改习题 4.12(a)中的公式,以便可以应用于不同 X 值的频数.

解 恰当的修改为

$$s = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2}$$

如习题 4.12(a), 由于

$$s = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}}$$

因此

$$\begin{aligned} s^2 &= \frac{\sum f(X - \bar{X})^2}{N} = \frac{\sum f(X^2 - 2\bar{X}X + \bar{X}^2)}{N} = \frac{\sum fX^2 - 2\bar{X} \sum fX + \bar{X}^2 \sum f}{N} \\ &= \frac{\sum fX^2}{N} - 2\bar{X} \frac{\sum fX}{N} + \bar{X}^2 = \frac{\sum fX^2}{N} - 2\bar{X}^2 + \bar{X}^2 = \frac{\sum fX^2}{N} - \bar{X}^2 \\ &= \frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2 \end{aligned}$$

或

$$s = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2}$$

注意在上述求和中,我们用 X 和 f 代替 X_j 和 f_j , \sum 代替 $\sum_{j=1}^K$, $\sum_{j=1}^K f_j = N$.

4.14 用习题 4.13 的公式求习题 4.11 表 4.2 中数据的标准差.

解 数据整理在表 4.3 中,其中 $\bar{X} = (\sum fX)/N = 67.45$ 英寸,注意这种方法像习题 4.11 那样会有复杂的计算.习题 4.17 所介绍的编码法可减少计算量.

表 4.3

身高(英寸)	组中值(X)	X^2	频数(f)	fX^2
60~62	61	3721	5	18 605
63~65	64	4096	18	73 728
66~68	67	4489	42	188 538
69~71	70	4900	27	132 300
72~74	73	5329	8	42 632
			$N = \sum f = 100$	$\sum fX^2 = 455 803$

$$s = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\frac{455 803}{100} - 67.45^2} = \sqrt{8.5275} = 2.92 \text{ 英寸}$$

4.15 如果 $d = X - A$ 表示 X 与任意常数 A 的差,证明

$$s = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

证明 由于 $d = X - A$, $X = A + d$, $\bar{X} = A + \bar{d}$ (见习题 3.18), 因此

$$X - \bar{X} = (A + d) - (A + \bar{d}) = d - \bar{d}$$

应用习题 4.13 的结论并用 d 和 \bar{d} 分别代表 X 和 \bar{X} ,

得到

$$s = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}} = \sqrt{\frac{\sum f(d - \bar{d})^2}{N}} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

另证

$$s^2 = \overline{(X - \bar{X})^2} = \overline{(d - \bar{d})^2} = \overline{d^2 - 2d\bar{d} + \bar{d}^2}$$

$$= \overline{d^2} - 2\bar{d}^2 + \bar{d}^2 = \overline{d^2} - \bar{d}^2 = \frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2$$

结果取正根.

- 4.16** 证明:如果在组距大小相等($=c$)的频数分布中,每个组中值 X 有关系 $X = A + cu$,其中 A 是某个给定组中值,那么标准差可写为

$$s = c \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N} \right)^2} = c \sqrt{\overline{u^2} - \bar{u}^2}$$

证明 由于 $d = X - A = cu$, c 是一常数,因此由习题 4.15,

$$\begin{aligned} s &= \sqrt{\frac{\sum f(cu)^2}{N} - \left(\frac{\sum f(cu)}{N} \right)^2} = \sqrt{c^2 \frac{\sum fu^2}{N} - c^2 \left(\frac{\sum fu}{N} \right)^2} \\ &= c \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N} \right)^2} \end{aligned}$$

另证 我们可不用习题 4.15 的结论而直接证明结果.由于 $X = A + cu$, $\bar{X} = A + c\bar{u}$, $X - \bar{X} = c(u - \bar{u})$, 于是

$$\begin{aligned} s^2 &= \overline{(X - \bar{X})^2} = \overline{c^2(u - \bar{u})^2} \\ &= c^2 \overline{(u^2 - 2\bar{u}u + \bar{u}^2)} = c^2(\overline{u^2} - 2\bar{u}^2 + \bar{u}^2) = c^2(\overline{u^2} - \bar{u}^2) \end{aligned}$$

得到

$$s = c \sqrt{\overline{u^2} - \bar{u}^2} = c \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N} \right)^2}$$

- 4.17** 根据(a)习题 4.15 的公式, (b)习题 4.16 的编码法求 XYZ 大学学生身高的标准差(见表 2.1).

解 (a)在表 4.4 和 4.5 中, A 是选定的组中值 67. 在表 4.4 中, $d = X - A$ 都是组距大小 $c = 3$ 的倍数. 在表 4.5 中消除了这个因素. 因此, 表 4.5 的计算量大大减少(与习题 4.11 和 4.14 相比较). 由于这个原因, 可尽可能的使用编码法.

表 4.4

组中值(X)	$d = X - A$	频数(f)	fd	fd^2
61	-6	5	-30	180
64	-3	18	-54	162
$A \rightarrow 67$	0	42	0	0
70	3	27	81	243
73	6	8	48	288
		$N = \sum f = 100$	$\sum fd = 45$	$\sum fd^2 = 873$

$$s = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2} = \sqrt{\frac{873}{100} - \left(\frac{45}{100} \right)^2} = \sqrt{8.5275} = 2.92 \text{ 英寸}$$

(b)见表 4.5.

表 4.5

组中值(X)	$u = \frac{X-A}{c}$	频数(f)	fu	fu^2
61	-2	5	-10	20
64	-1	18	-18	18
A→67	0	42	0	0
70	1	27	27	27
73	2	8	16	32
		$N = \sum f = 100$	$\sum fu = 15$	$\sum fu^2 = 97$

$$s = c \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 3 \sqrt{\frac{97}{100} - \left(\frac{15}{100}\right)^2} = 3 \sqrt{0.9475} = 2.92 \text{ 英寸}$$

4.18 用编码法求 P&R 公司 65 个员工工资分布的(a)均值,(b)标准差(见习题 2.3 表 2.4).

解 数据整理在表 4.6 中.

表 4.6

X(美元)	u	f	fu	fu ²
255.0	-2	8	-16	32
265.00	-1	10	-10	10
A→275.00	0	16	0	0
285.00	1	14	14	14
~ 295.00	2	10	20	40
305.00	3	5	15	45
315.00	4	2	8	32
		$N = \sum f = 65$	$\sum fu = 31$	$\sum fu^2 = 173$

$$(a) \bar{X} = A + c\bar{u} = A + c \frac{\sum fu}{N} = 275.00 + 10.00 \times \frac{31}{65} = 279.77 \text{ 美元}$$

$$(b) s = c \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 10.00 \sqrt{\frac{173}{65} - \left(\frac{31}{65}\right)^2} \\ = 10.00 \sqrt{2.4341} = 15.60 \text{ 美元}$$

4.19 表 4.7 显示了某小学 480 个学生的 IQ. 用编码法求(a)均值,(b)标准差.

表 4.7

组中值(X)	70	74	78	82	86	90	94	98	102	106	110	114	118	122	126
频数(f)	4	9	16	28	45	66	85	72	54	38	27	18	11	5	2

解 智商

$$IQ = \frac{\text{智力年龄}}{\text{年龄}}$$

用百分数表示. 例如, 一个 8 岁儿童(接受某种教育)的智力相当于一个 10 岁儿童的智力, 相应的 IQ 是 $10/8 = 1.25 = 125\%$ 或简单记为 125, % 符号不言自明.

为了求表 4.7 中 IQ 的均值和标准差, 我们把数据整理在表 4.8 中.

$$(a) \bar{X} = A + c\bar{u} = A + c \frac{\sum fu}{N} = 94 + 4 \times \frac{236}{480} = 95.97$$

(b)

$$\begin{aligned} s &= c \sqrt{u^2 - \bar{u}^2} = c \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} \\ &= 4 \sqrt{\frac{3404}{480} - \left(\frac{236}{480}\right)^2} = 4 \sqrt{6.8499} = 10.47 \end{aligned}$$

表 4.8

X	u	f	fu	fu ²
70	-6	4	-24	144
74	-5	9	-45	225
78	-4	16	-64	256
82	-3	28	-84	252
86	-2	45	-90	180
90	-1	66	-66	66
A → 94	0	85	0	0
98	1	72	72	72
102	2	54	108	216
106	3	38	114	342
110	4	27	108	432
114	5	18	90	450
118	6	11	66	396
122	7	5	35	245
126	8	2	16	128
		$N = \sum f = 480$	$\sum fu = 236$	$\sum fu^2 = 3404$

Charlier 检验

4.20 运用 Charlier 检验来帮助证实习题 4.19 中(a)均值,(b)标准差的计算.

解 为了运用 Charlier 检验,表 4.9 由表 4.8 而得(为了方便,第二列照写).

表 4.9

u + 1	f	f(u + 1)	f(u + 1) ²
-5	4	-20	100
-4	9	-36	144
-3	16	-48	144
-2	28	-56	112
-1	45	-45	45
0	66	0	0
1	85	85	85
2	72	144	288
3	54	162	486
4	38	152	608
5	27	135	675
6	18	108	648
7	11	77	539
8	5	40	320
9	2	18	162
$N = \sum f = 480$		$\sum f(u + 1) = 716$	$\sum f(u + 1)^2 = 4356$

- (a) 由表 4.9, $\sum f(u+1) = 716$; 由表 4.8, $\sum fu + N = 236 + 480 = 716$. 这是均值的检验.
 (b) 由表 4.9, $\sum f(u+1)^2 = 4356$, 由表 4.8, $\sum fu^2 + 2\sum fu + N = 3404 + 2 \times 236 + 480 = 4356$.
 这是标准差的检验.

Sheppard 方差修正

4.21 应用 Sheppard 修正求(a)习题 4.17, (b)习题 4.18, (c)习题 4.19 的标准差.

- 解 (a) $s^2 = 8.5275$, $c = 3$. 修正方差 $= s^2 - c^2/12 = 8.5275 - 3^2/12 = 7.7775$. 修正标准差 $= \sqrt{\text{修正方差}} = \sqrt{7.7775} = 2.79$ 英寸.
 (b) $s^2 = 243.41$, $c = 10$. 修正方差 $= s^2 - c^2/12 = 243.41 - 10^2/12 = 235.08$. 修正标准差 $= \sqrt{\text{修正方差}} = \sqrt{235.08} = 15.33$ 美元.
 (c) $s^2 = 109.60$, $c = 4$. 修正方差 $= s^2 - c^2/12 = 109.60 - 4^2/12 = 108.27$. 修正标准差 $= \sqrt{\text{修正方差}} = \sqrt{108.27} = 10.41$.

4.22 求习题 2.8 第二种频数分布(见表 2.7)的(a)均值, (b)标准差, (c)Sheppard 修正标准差, (d)根据原始数据得到的真实标准差.

解 数据整理在表 4.10 中.

表 4.10

X	u	f	fu	fu^2
122	-3	3	-9	27
131	-2	5	-10	20
140	-1	9	-9	9
$A \rightarrow 149$	0	12	0	0
158	1	5	5	5
167	2	4	8	16
176	3	2	6	18
		$N = \sum f = 40$	$\sum fu = -9$	$\sum fu^2 = 95$

$$(a) \bar{X} = A + c\bar{u} = A + c \frac{\sum fu}{N} = 149 + 9 \times \frac{-9}{40} = 147.0 \text{ 磅}$$

(b)

$$s = c \sqrt{u^2 - \bar{u}^2} = c \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2}$$

$$= 9 \sqrt{\frac{95}{40} - \left(\frac{-9}{40}\right)^2} = 9 \sqrt{2.324375} = 13.7 \text{ 磅}$$

(c) 修正方差 $= s^2 - c^2/12 = 188.27 - 9^2/12 = 181.52$. 修正标准差 $= 13.5$ 磅.

(d) 为了从实际数据中求得标准差, 最好从每个数据中减去一个恰当的值 A , 比如 $A = 150$, 然后运用习题 4.15 中的方法. $d = X - A = X - 150$ 列在下表中:

-12	14	0	-18	-6	-25	-1	7
-4	8	-10	-3	-14	-2	2	-6
18	-24	-12	26	13	-31	4	15
-4	23	-8	-3	-15	3	-10	-15
11	-5	-15	-8	0	6	-5	-22

从中我们得到 $\sum d = -128$, $\sum d^2 = 7052$.

$$s = \sqrt{d^2 - \bar{d}^2} = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} = \sqrt{\frac{7052}{40} - \left(\frac{-128}{40}\right)^2} = \sqrt{166.06} = 12.9 \text{ 磅}$$

因此, Sheppard 修正在这里提供了一些改进.

离差度量间的经验关系

- 4.23 根据 XYZ 大学学生身高的分布, 讨论下列经验公式的合理性: (a) 平均偏差 = $\frac{4}{5}$ 标准差, (b) 半内四分位数间距 = $\frac{2}{3}$ 标准差.

解 (a) 由习题 4.4 和 4.11, 平均偏差 ÷ 标准差 = $2.26/2.92 = 0.77$, 接近于 $\frac{4}{5}$.

(b) 由习题 4.6 和 4.11, 半内四分位数间距 ÷ 标准差 = $1.98/2.92 = 0.68$, 接近于 $\frac{2}{3}$.

因此, 经验公式在本例中是合理的.

由于没有相应的平均偏差或半内四分位数间距的修正, 在上面我们没有使用 Sheppard 修正标准差.

标准差的性质

- 4.24 确定习题 4.19 中学生 IQ 落在区域 (a) $\bar{X} \pm s$; (b) $\bar{X} \pm 2s$; (c) $\bar{X} \pm 3s$ 中的百分比.

解 (a) $\bar{X} \pm s = 95.97 \pm 10.47$, 即区域是 85.5 ~ 106.4. IQ 在区域 $\bar{X} \pm s$ 中的数目为

$$\frac{88 - 85.5}{4} \times 45 + 66 + 85 + 72 + 54 + \frac{106.4 - 104}{4} \times 38 = 339$$

IQ 落在区域 $\bar{X} \pm s$ 中的百分比是 $339/480 = 70.6\%$.

(b) $\bar{X} \pm 2s = 95.97 \pm 2 \times 10.47$, 即区域是 75.0 ~ 116.9. IQ 在区域 $\bar{X} \pm 2s$ 中的数目为

$$\begin{aligned} \frac{76 - 75.0}{4} \times 9 + 16 + 28 + 45 + 66 + 85 + 72 + 54 \\ + 38 + 27 + 18 + \frac{116.9 - 116}{4} \times 11 = 451 \end{aligned}$$

IQ 落在区域 $\bar{X} \pm 2s$ 中的百分比是 $451/480 = 94.0\%$.

(c) $\bar{X} \pm 3s = 95.97 \pm 3 \times 10.47$, 即区域是 64.6 ~ 127.4. IQ 在区域 $\bar{X} \pm 3s$ 中的数目为

$$480 - \frac{128 - 127.4}{4} \times 2 = 479.7 \text{ 或 } 480$$

IQ 落在区域 $\bar{X} \pm 3s$ 中的百分比是 $479.7/480 = 99.9\%$, 或 100% .

(a), (b), (c) 中的百分比分别与正态分布对应的百分比 68.27%, 95.45%, 99.73% 是一致的.

这里我们没有使用 Sheppard 修正标准差, 应用 Sheppard 修正的结果与上述结果相当接近. 注意上述结果可由习题 4.32 表 4.11 得到.

- 4.25 给定数列 2, 5, 8, 11, 14 和 2, 8, 14. 求: (a) 每组的均值, (b) 每组的方差, (c) 两组联合的均值, (d) 两组联合的方差.

解 (a) 第一组均值 = $\frac{1}{5}(2 + 5 + 8 + 11 + 14) = 8$. 第二组均值 = $\frac{1}{3}(2 + 8 + 14) = 8$.

(b) 第一组方差 = $s_1^2 = \frac{1}{5}[(2-8)^2 + (5-8)^2 + (8-8)^2 + (11-8)^2 + (14-8)^2] = 18$.

第二组方差 = $s_2^2 = \frac{1}{3}[(2-8)^2 + (8-8)^2 + (14-8)^2] = 24$.

(c) 两组联合的均值

$$\frac{2 + 5 + 8 + 11 + 14 + 2 + 8 + 14}{5 + 3} = 8$$

(d) 两组联合的方差

$$\begin{aligned} s^2 &= [(2-8)^2 + (5-8)^2 + (8-8)^2 + (11-8)^2 \\ &\quad + (14-8)^2 + (2-8)^2 + (8-8)^2 + (14-8)^2] / (5 + 3) \\ &= 20.25 \end{aligned}$$

另解 使用公式

$$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2} = \frac{5 \times 18 + 3 \times 24}{5 + 3} = 20.25$$

- 4.26 给定数列 2, 5, 8, 11, 14 和 10, 16, 22, 回答习题 4.25 中的提问.

解 两数列的均值分别为 8 和 16, 方差分别为 $s_1^2 = 18, s_2^2 = 24$.

$$\text{联合的均值} = \frac{2+5+8+11+14+10+16+22}{5+3} = 11$$

联合的方差

$$\begin{aligned} s^2 &= [(2-11)^2 + (5-11)^2 + (8-11)^2 + (11-11)^2 + (14-11)^2 \\ &\quad + (10-11)^2 + (16-11)^2 + (22-11)^2] / (5+3) \\ &= 35.25 \end{aligned}$$

注意, 由于两数列均值不同, 因此公式 $s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2}$ 在这里不能采用.

4.27 (a) 证明 $w^2 + pw + q$ 当且仅当 $w = -\frac{1}{2}p$ 时, 取得最小值, 其中 p, q 是给定常数.

(b) 根据(a), 证明当且仅当 $a = \bar{X}$ 时,

$$\frac{\sum_{i=1}^N (X_i - a)^2}{N} \text{ 或简记为 } \frac{\sum (X - a)^2}{N}$$

取得最小值.

证明 (a) 我们知道 $w^2 + pw + q = \left(w + \frac{1}{2}p\right)^2 + q - \frac{1}{4}p^2$. 由于 $q - \frac{1}{4}p^2$ 是常数, 表达式当且仅当 $w + \frac{1}{2}p = 0$ (即 $w = -\frac{1}{2}p$) 时, 取得最小值.

(b)

$$\begin{aligned} \frac{\sum (X - a)^2}{N} &= \frac{\sum (X^2 - 2aX + a^2)}{N} = \frac{\sum X^2 - 2a \sum X + Na^2}{N} \\ &= a^2 - 2a \frac{\sum X}{N} + \frac{\sum X^2}{N} \end{aligned}$$

把最后一项与 $w^2 + pw + q$ 作比较, 我们得到

$$w = a, \quad p = -2 \frac{\sum X}{N}, \quad q = \frac{\sum X^2}{N}$$

根据 (a) 的结论, 当 $a = -\frac{1}{2}p = (\sum X)/N = \bar{X}$ 时, 表达式取得最小值.

绝对和相对离差, 变异系数

4.28 电视显像管生产商制造两种显像管 A 和 B. 它们的平均寿命分别为 $\bar{X}_A = 1495$ 小时, $\bar{X}_B = 1875$ 小时, 标准差为 $s_A = 280$ 小时, $s_B = 310$ 小时. 请问哪种显像管有较大的 (a) 绝对离差, (b) 相对离差.

解 (a) A 的绝对离差 $s_A = 280$ 小时, B 的绝对离差 $s_B = 310$ 小时. 因此 B 的绝对离差比 A 的大.

(b) 变异系数

$$A = \frac{s_A}{\bar{X}_A} = \frac{280}{1495} = 18.7\%, \quad B = \frac{s_B}{\bar{X}_B} = \frac{310}{1875} = 16.5\%$$

因此 A 的相对离差较大.

4.29 用未修正和修正标准差求 (a) 习题 4.14; (b) 习题 4.18 中数据的变异系数.

解 (a) $V(\text{未修正}) = \frac{s(\text{未修正})}{\bar{X}} = \frac{2.92}{67.45} = 0.0433 = 4.3\%$

$$V(\text{修正}) = \frac{s(\text{修正})}{\bar{X}} = \frac{2.79}{67.45} = 0.0413 = 4.1\% \quad (\text{根据习题 4.21(a)})$$

(b) $V(\text{未修正}) = \frac{s(\text{未修正})}{\bar{X}} = \frac{15.60}{79.77} = 0.196 = 19.6\%$

$$V(\text{修正}) = \frac{s(\text{修正})}{\bar{X}} = \frac{15.33}{79.77} = 0.192 = 19.2\% \quad (\text{根据习题 4.21(b)})$$

4.30 (a)若四分位数已知,定义一组数据相对离差的度量.

(b)用习题 4.6 中数据来说明(a)中定义度量的计算.

解 (a) 如果数据的 Q_1 和 Q_3 给定,那么 $\frac{1}{2}(Q_1 + Q_3)$ 是数据集中趋势(或平均值)的度量,而半内四分位数间距 $Q = \frac{1}{2}(Q_3 - Q_1)$ 度量的是数据的离差.因此我们可定义一个相对离差的度量:

$$V_Q = \frac{\frac{1}{2}(Q_3 - Q_1)}{\frac{1}{2}(Q_1 + Q_3)} = \frac{Q_3 - Q_1}{Q_1 + Q_3}$$

我们称它为四分位变异系数或四分位相对离差系数.

$$(b) V_Q = \frac{Q_3 - Q_1}{Q_1 + Q_3} = \frac{69.61 - 65.64}{69.61 + 65.64} = \frac{3.97}{135.25} = 0.0293 = 2.9\%$$

标准化变量,标准分数

4.31 一个学生数学期末成绩是 84,该门功课成绩的平均分是 76,标准差是 10.她物理期末成绩是 90,该门功课成绩的平均分是 82,标准差是 16.她在哪门功课中名次更前?

解 标准化变量 $z = (X - \bar{X})/s$ 是根据标准差 s 来度量 X 与均值 \bar{X} 之间的偏差.对于数学, $z = (84 - 76)/10 = 0.8$;对于物理, $z = (90 - 82)/16 = 0.5$.因此,这个学生的数学成绩高于平均分 0.8 个标准分,而物理的标准离差高于平均分 0.5,所以她数学名次更靠前.

变量 $z = (X - \bar{X})/s$ 常用来进行教育测评,也就是众所周知的标准分数.

4.32 (a)把习题 4.19 中的 IQ 转化为标准分数;(b)相对于标准分数,建立频率图.

解 (a) 转化工作在表 4.11 中进行.

IQ 组中值 66, 130 放入表中是为了在(b)中使用.这里没有使用 Sheppard 修正,这个问题的修正得分与表 4.11 所示的一样.

(b)相对于 z 分数的频率图(频率多边形)如图 4-2 所示.水平轴把标准差 s 作为一个单位.注意这里的分布是微对称的,稍向右斜.

表 4.11 $\bar{X} = 96.0, s = 10.5$

IQ(X)	$X - \bar{X}$	$z = \frac{X - \bar{X}}{s}$	频数(f)	频率 (f)/N(%)
66	-30.0	-2.86	0	0.0
70	-26.0	-2.48	4	0.8
74	-22.0	-2.10	9	1.9
78	-18.0	-1.71	16	3.3
82	-14.0	-1.33	28	5.8
86	-10.0	-0.95	45	9.4
90	-6.0	-0.57	66	13.8
94	-2.0	-0.19	85	17.7
98	2.0	0.19	72	15.0
102	6.0	0.57	54	11.2
106	10.0	0.95	38	7.9
110	14.0	1.33	27	5.6
114	18.0	1.71	18	3.8
118	22.0	2.10	11	2.3
122	26.0	2.48	5	1.0
126	30.0	2.86	2	0.4
130	34.0	3.24	0	0.0
			480	100

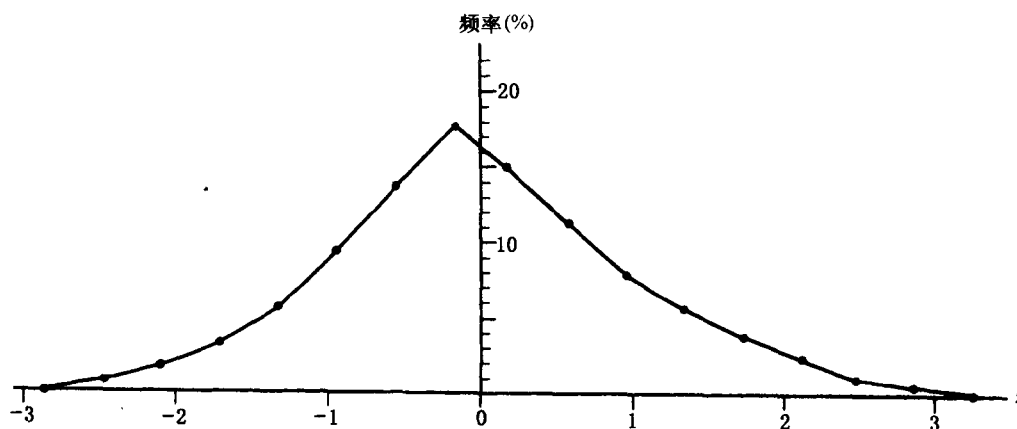


图 4-2

补充习题

全距

- 4.33 求(a)5, 3, 8, 4, 7, 6, 12, 4, 3 和(b)8.772, 6.453, 10.624, 8.628, 9.434, 6.351 的全距.
- 4.34 根据习题 3.59 的表 3.8, 求最大负荷量的全距.
- 4.35 根据习题 3.61 的表 3.10, 求铆钉直径的全距.
- 4.36 50 个测量值的最大值是 8.34 千克. 如果全距是 0.46 千克, 求测量值的最小值.
- 4.37 下面的表格给出了 25 个由于公司裁员而失业的老工人寻找工作所需要的时间(星期). 求这些数据的全距.

13	13	17	7	22
22	26	17	13	14
16	7	6	18	20
10	17	11	10	15
16	8	16	21	11

平均偏差

- 4.38 求(a) -18.2 ; (b) $+3.58$; (c) 6.21 ; (d) 0 ; (e) $-\sqrt{2}$; (f) $4.00 - 2.36 - 3.52$ 的绝对值.
- 4.39 求(a)3, 7, 9, 5; (b)2.4, 1.6, 3.8, 4.1, 3.4 的平均偏差.
- 4.40 求习题 4.33 中数据的平均偏差.
- 4.41 根据习题 3.59 表 3.8, 求最大负荷量的平均偏差.
- 4.42 (a)根据习题 3.61 表 3.10, 计算铆钉直径的平均偏差(MD).
(b)求铆钉直径落在区域 $(\bar{X} \pm MD)$, $(\bar{X} \pm 2MD)$ 和 $(\bar{X} \pm 3MD)$ 之间的百分率.
- 4.43 根据(a)均值, (b)中位数, 求 8, 10, 9, 12, 4, 8, 2 的平均偏差. 证明根据中位数而求得平均偏差不大于根据均值而求得平均偏差.
- 4.44 根据习题 3.60 表 3.9 的分布, 求(a)关于均值, (b)关于中位数的平均偏差. 利用习题 3.60 和 3.70 的结论.
- 4.45 根据习题 3.62 表 3.11 的分布, 求(a)关于均值, (b)关于中位数的平均偏差. 利用习题 3.62 和 3.72 的结论.
- 4.46 根据习题 4.37 中的数据, 求平均偏差.

- 4.47 写出计算(a)关于均值, (b)关于中位数的平均偏差的编码公式. 应用这些公式证明习题 4.44 和 4.45 的结论.

半内四分位数间距

- 4.48 根据(a)习题 3.59, (b)习题 3.60, (c)习题 3.107 中的分布, 求半内四分位数间距.
- 4.49 根据习题 4.37 所给数据求半内四分位数间距.
- 4.50 证明在任何频数分布中落在区间 $\frac{1}{2}(Q_1 + Q_3) \pm \frac{1}{2}(Q_3 - Q_1)$ 内的数据为 50%. 那么区间 $Q_2 \pm \frac{1}{2}(Q_3 - Q_1)$ 呢? 证明你的结论.
- 4.51 (a)如何根据频数分布画出相应的半内四分位数间距?
(b)请解释半内四分位数间距与卵形线之间的关系.

10~90 百分位数间距

- 4.52 根据(a)习题 3.59, (b)习题 3.107 中的分布, 计算 10~90 百分位数间距, 并解释你的结论.
- 4.53 一城市家庭购买支出的第 10 个百分位数是 35500 美元, 第 90 个百分位数是 225 000 美元. 求 10~90 百分位数间距并给出包含 80% 购买支出的间距.
- 4.54 与 10~90 百分位数间距相比, 20~80 百分位数间距有哪些优势和劣势?
- 4.55 用(a)10~90 百分位数间距, (b)20~80 百分位数间距, (c) 25~75 百分位数间距解答习题 4.51. 并回答(c)与半内四分位数间距之间的关系.

标准差

- 4.56 求(a)3, 6, 2, 1, 7, 5; (b)3.2, 4.6, 2.8, 5.2, 4.4; (c)0, 0, 0, 0, 0, 1, 1, 1 的标准差.
- 4.57 (a)在数组 3, 6, 2, 1, 7, 5 中的每个数字上加 5, 得到 8, 11, 7, 6, 12, 10. 证明两组数有相同的标准差和不同的均值, 并说出均值之间的关系.
(b)数组 3, 6, 2, 1, 7, 5 中的每个数字乘以 2 再加上 5, 得到 11, 17, 9, 7, 19, 15. 请说出两组数标准差和均值之间的关系.
(c)通过具体问题(a)和(b)揭示了均值和标准差的什么性质?
- 4.58 求等差数列 4, 10, 16, 22, ..., 154 的标准差.
- 4.59 求(a)习题 3.59, (b)习题 3.60, (c)习题 3.107 中分布的标准差.
- 4.60 在习题 4.59 的每个部分中, 用实例说明 Charlier 检验的用途.
- 4.61 求习题 2.17 分布的(a)均值, (b)标准差, 并解释所得结果的含义.
- 4.62 如果数据服从钟形分布, 那么标准差可近似表示为全距的四分之一. 计算习题 4.37 所给数据的标准差, 并与全距的四分之一作比较.
- 4.63 (a)根据习题 3.61 表 3.10, 求铆钉直径的标准差 s .
(b)铆钉直径落入区域 $\bar{X} \pm s$, $\bar{X} \pm 2s$ 和 $\bar{X} \pm 3s$ 中的百分比是多少?
(c)如果数据服从正态分布, 把(b)中的百分比与理论上推出的结果作比较, 并解释其中的差异.
- 4.64 应用 Sheppard 修正求习题 4.59 中每部分的标准差, 并讨论在每种情况下, 应用此方法是否恰当.
- 4.65 如果应用 Sheppard 修正, 那么习题 4.63 的结论会作如何调整?
- 4.66 (a)求习题 2.8 中数据的均值和标准差.
(b)建立数据的频数分布, 并求标准差.
(c)比较(a)和(b)中的结果, 讨论应用 Sheppard 修正后是否会得到较好的结论.
- 4.67 根据习题 2.27 中的数据解答习题 4.66.
- 4.68 (a)有 N 个取 1 或 0 的数, 其中取“1”与取“0”的个数之比为 $p:q$ ($q=1-p$). 证明这组数的标准差为 \sqrt{pq} .
(b)用(a)中结论解答习题 4.56(c).
- 4.69 (a)证明 n 个数 $a, a+d, a+2d, \dots, a+(n-1)d$ (即首项为 a , 公差为 d 的等差数列) 的方差为 $\frac{1}{12}(n^2-1)d^2$.

(b)用(a)中的结论解答习题 4.58.(提示: $1+2+3+\cdots+(n-1)=\frac{1}{2}n(n-1)$, $1^2+2^2+3^2+\cdots+(n-1)^2=\frac{1}{6}n(n-1)(2n-1)$).

4.70 推广并证明本章性质 3, 即(12)式.

离差度量间的经验关系

4.71 通过比较习题 4.59 得到的标准差与习题 4.41, 4.42 和 4.44 对应的平均偏差, 讨论下列经验关系是否成立: 平均偏差 $= \frac{4}{5}$ 标准差. 解释可能出现的差异.

4.72 通过比较习题 4.59 得到的标准差与习题 4.48 对应的半内四分位数间距, 讨论下列经验关系是否成立: 半内四分位数间距 $= \frac{2}{3}$ 标准差. 解释可能出现的差异.

4.73 你认为对于微斜的钟形分布, 半内四分位数间距和平均偏差之间存在何种经验关系?

4.74 近似于正态分布的频数分布的半内四分位数间距等于 10. 那么(a)标准差, (b)平均偏差是多少?

绝对和相对离差; 变异系数

4.75 150 个学生统计学期末考试平均成绩为 78, 标准差是 8.0, 代数学期末考试平均成绩为 73, 标准差是 7.6. 哪一门学科的(a)绝对离差, (b)相对离差更大?

4.76 求(a)习题 3.59, (b)习题 3.107 中数据的变异系数.

4.77 一群中学生 SAT 成绩分布的第一个四分位数是 825, 第三个四分位数是 1125. 求这些学生 SAT 成绩分布的四分位变异系数.

4.78 对于 15~24 年龄组, 家庭收入的第一个四分位数是 16500 美元, 第三个四分位数是 25 000 美元. 求这个年龄组家庭收入的四分位变异系数.

标准化变量, 标准分数

4.79 在习题 4.75 中, 一个学生统计学得 75 分, 代数学得 71 分. 请问他在哪门学科名次较前?

4.80 把 6, 2, 8, 7, 5 转化为标准分数.

4.81 证明一组标准分数的均值和方差分别为 0 和 1. 用习题 4.80 来说明.

4.82 (a)把习题 3.107 的成绩转化为标准分数, (b)相对于标准分数建立频率曲线.

第五章 矩, 偏度和峰度

矩

如果 X_1, X_2, \dots, X_N 是变量 X 的 N 个值, 定义

$$\overline{X^r} = \frac{X_1^r + X_2^r + \dots + X_N^r}{N} = \frac{\sum_{j=1}^N X_j^r}{N} = \frac{\sum X^r}{N} \quad (1)$$

为 r 阶矩. 当 $r=1$ 时, 一阶矩就是均值 \bar{X} .

r 阶中心矩定义为

$$m_r = \frac{\sum_{j=1}^N (X_j - \bar{X})^r}{N} = \frac{\sum (X - \bar{X})^r}{N} = \overline{(X - \bar{X})^r} \quad (2)$$

当 $r=1$ 时, $m_1=0$ (见习题 3.16). 当 $r=2$ 时, $m_2=s^2$ 为方差.

设 A 为任意给定的一个数, 关于原点 A 的 r 阶矩定义为

$$m'_r = \frac{\sum_{j=1}^N (X_j - A)^r}{N} = \frac{\sum (X - A)^r}{N} = \frac{\sum d^r}{N} = \overline{(X - A)^r} \quad (3)$$

其中 $d = X - A$ 是 X 与 A 的差. 如果 $A=0$, (3) 式即为 (1) 式. 因为这个原因, (1) 式也被称为 r 阶原点矩.

分类资料的矩

如果 X_1, X_2, \dots, X_K 发生的频数分别为 f_1, f_2, \dots, f_K . 上述矩可写为

$$\overline{X^r} = \frac{f_1 X_1^r + f_2 X_2^r + \dots + f_K X_K^r}{N} = \frac{\sum_{j=1}^K f_j X_j^r}{N} = \frac{\sum f X^r}{N} \quad (4)$$

$$m_r = \frac{\sum_{j=1}^K f_j (X_j - \bar{X})^r}{N} = \frac{\sum f (X - \bar{X})^r}{N} = \overline{(X - \bar{X})^r} \quad (5)$$

$$m'_r = \frac{\sum_{j=1}^K f_j (X_j - A)^r}{N} = \frac{\sum f (X - A)^r}{N} = \overline{(X - A)^r} \quad (6)$$

其中 $N = \sum_{j=1}^K f_j = \sum f$. 对于分类资料的矩计算, 上述公式是很有益处的.

矩间关系

下列公式对 m_r 和 m'_r 成立:

$$\begin{aligned} m_2 &= m'_2 - m_1'^2 \\ m_3 &= m'_3 - 3m_1' m_2' + 2m_1'^3 \\ m_4 &= m'_4 - 4m_1' m_3' + 6m_1'^2 m_2' - 3m_1'^4 \end{aligned} \quad (7)$$

等等 (见习题 5.5). 注意, $m_1' = \bar{X} - A$.

分类资料矩的计算

前面几章介绍过用来计算均值和标准差的编码法是矩计算的短方法. 由于 $X_j = A + cu_j$,

(或简记为 $X = A + cu$), 因此由(6)式, 我们可知

$$m'_r = c^r \frac{\sum fu^r}{N} = c^r \overline{u^r} \quad (8)$$

再由(7)式, 就可求得 m_r .

Charlier 检验和 Sheppard 修正

用编码法计算矩的 Charlier 检验运用了以下性质:

$$\begin{aligned} \sum f(u+1) &= \sum fu + N \\ \sum f(u+1)^2 &= \sum fu^2 + 2 \sum fu + N \\ \sum f(u+1)^3 &= \sum fu^3 + 3 \sum fu^2 + 3 \sum fu + N \\ \sum f(u+1)^4 &= \sum fu^4 + 4 \sum fu^3 + 6 \sum fu^2 + 4 \sum fu + N \end{aligned} \quad (9)$$

Sheppard 矩修正(第 71 页上思路的延伸)如下所示:

$$\text{修正的 } m_2 = m_2 - \frac{1}{12}c^2, \text{ 修正的 } m_4 = m_4 - \frac{1}{2}c^2m_2 + \frac{7}{240}c^4$$

矩 m_1 和 m_3 无须修正.

无量纲形式的矩

为了避免特殊的单位, 我们定义无量纲矩:

$$a_r = \frac{m_r}{s^r} = \frac{m_r}{(\sqrt{m_2})^r} = \frac{m_r}{\sqrt{m_2^r}} \quad (10)$$

其中 $s = \sqrt{m_2}$ 是标准差. 由于 $m_1 = 0$, $m_2 = s^2$, 我们得到 $a_1 = 0$, $a_2 = 1$.

偏度

偏度是一个分布中不对称程度或偏离对称程度的反映. 如果分布的频数曲线(光滑频数多边形)右边的尾部比左边的长, 则称分布是**向右偏**的或有**正偏度**. 反之, 则称分布是**向左偏**的或**负偏度**.

对于斜分布, 均值和众数都落在尾部较长的一边(见图 3-1 和 3-2). 因此, 均值与众数的差就可用来度量不对称性. 如果再除以离差, 比如标准差, 就可得到偏度的无量纲形式:

$$\text{偏度} = \frac{\text{均值} - \text{众数}}{\text{标准差}} = \frac{\bar{X} - \text{众数}}{s} \quad (11)$$

如不用众数, 可以用第三章经验公式(10):

$$\text{偏度} = \frac{3(\text{均值} - \text{中位数})}{\text{标准差}} = \frac{3(\bar{X} - \text{中位数})}{s} \quad (12)$$

(11)和(12)式分别称为 **Pearson 第一、第二偏度系数**.

根据四分位数和百分位数可定义其他偏度量:

$$\text{四分位偏度系数} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} \quad (13)$$

$$10 \sim 90 \text{ 百分位偏度系数} = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{P_{90} - P_{10}} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} \quad (14)$$

也可用无量纲形式的三阶矩来度量偏度:

$$\text{矩偏度系数} = a_3 = \frac{m_3}{s^3} = \frac{m_3}{(\sqrt{m_2})^3} = \frac{m_3}{\sqrt{m_2^3}} \quad (15)$$

有时也用 $b_1 = a_3^2$ 来度量偏度. 对完全对称的曲线, 比如正态曲线, a_3 和 b_1 都是 0.

峰度

峰度是分布陡峭程度的反映,通常是相对于正态分布而言.有一个相对较高的顶峰的分布,如图 5-1(a),称为**尖峰**的,而图 5-1(b)顶峰较为平坦,称为**扁峰**的.图 5-1(c)所示的正态分布,没有较高和较平坦的顶峰,称为**常峰态**的.

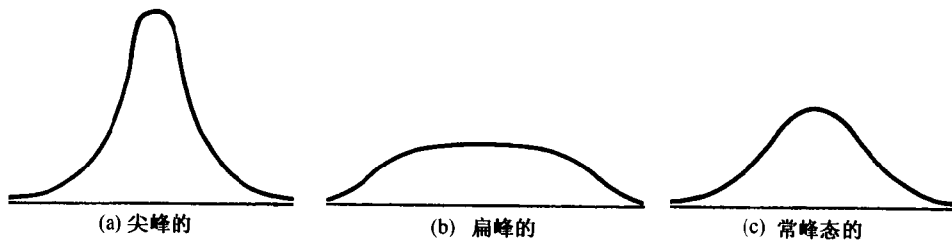


图 5-1

表示为无量纲形式的四阶中心矩可用来度量峰度:

$$\text{矩峰度系数} = a_4 = \frac{m_4}{s^4} = \frac{m_4}{m_2^2} \quad (16)$$

a_4 也常记作 b_2 . 对于正态分布, $b_2 = a_4 = 3$. 因此,峰度有时也定义为 $(b_2 - 3)$. $(b_2 - 3)$ 取正的就意味着尖峰分布,取负的就意味着扁峰分布,值为 0 就是正态分布.

另一种度量峰度的方法建立在四分位数和百分位数的基础上,定义为

$$\kappa = \frac{Q}{P_{90} - P_{10}} \quad (17)$$

其中 $Q = \frac{1}{2}(Q_3 - Q_1)$ 是半内四分位数间距. 我们把 κ 称为**百分位峰度系数**;对于正态分布, $\kappa = 0.263$ (见习题 5.14).

总体矩, 偏度和峰度

如果需要区别样本与总体的矩, 偏度和峰度, 那么习惯上用拉丁符号表示前者, 用希腊符号表示后者. 因此, 如果样本矩表示为 m_r, m'_r , 则相应的总体矩表示为 μ_r, μ'_r . 下标通常用拉丁符号表示.

同样的, 如果样本偏度和峰度分别表示为 a_3, a_4 , 则总体偏度和峰度分别为 α_3, α_4 .

在第四章我们已经知道样本的标准差和总体的标准差分别表示为 s 和 σ .

习题及解答

5.1 求数集 2, 3, 7, 8, 10 的 (a) 一阶矩, (b) 二阶矩, (c) 三阶矩, (d) 四阶矩.

解 (a) 一阶矩, 或均值为

$$\bar{X} = \frac{\sum X}{N} = \frac{2+3+7+8+10}{5} = \frac{30}{5} = 6$$

(b) 二阶矩为

$$\overline{X^2} = \frac{\sum X^2}{N} = \frac{2^2+3^2+7^2+8^2+10^2}{5} = \frac{226}{5} = 45.2$$

(c) 三阶矩为

$$\overline{X^3} = \frac{\sum X^3}{N} = \frac{2^3+3^3+7^3+8^3+10^3}{5} = \frac{1890}{5} = 378$$

(d) 四阶矩为

$$\overline{X^4} = \frac{\sum X^4}{N} = \frac{2^4 + 3^4 + 7^4 + 8^4 + 10^4}{5} = \frac{16\,594}{5} = 3318.8$$

5.2 求习题 5.1 中数据的 (a) 一阶中心矩, (b) 二阶中心矩, (c) 三阶中心矩, (d) 四阶中心矩.

解 (a)

$$\begin{aligned} m_1 &= \overline{(X - \bar{X})} = \frac{\sum (X - \bar{X})}{N} \\ &= \frac{(2-6) + (3-6) + (7-6) + (8-6) + (10-6)}{5} = \frac{0}{5} = 0 \end{aligned}$$

由于 $\overline{X - \bar{X}} = \bar{X} - \bar{X} = 0$ (见习题 3.16), $m_1 = 0$.

(b)

$$\begin{aligned} m_2 &= \overline{(X - \bar{X})^2} = \frac{\sum (X - \bar{X})^2}{N} \\ &= \frac{(2-6)^2 + (3-6)^2 + (7-6)^2 + (8-6)^2 + (10-6)^2}{5} = \frac{46}{5} = 9.2 \end{aligned}$$

注意, m_2 就是方差 s^2 .

(c)

$$\begin{aligned} m_3 &= \overline{(X - \bar{X})^3} = \frac{\sum (X - \bar{X})^3}{N} \\ &= \frac{(2-6)^3 + (3-6)^3 + (7-6)^3 + (8-6)^3 + (10-6)^3}{5} = \frac{-18}{5} = -3.6 \end{aligned}$$

(d)

$$\begin{aligned} m_4 &= \overline{(X - \bar{X})^4} = \frac{\sum (X - \bar{X})^4}{N} \\ &= \frac{(2-6)^4 + (3-6)^4 + (7-6)^4 + (8-6)^4 + (10-6)^4}{5} = \frac{610}{5} = 122 \end{aligned}$$

5.3 求习题 5.1 中数据的关于 4 的原点 (a) 一阶矩, (b) 二阶矩, (c) 三阶矩, (d) 四阶矩.

解 (a)

$$\begin{aligned} m'_1 &= \overline{(X - 4)} = \frac{\sum (X - 4)}{N} \\ &= \frac{(2-4) + (3-4) + (7-4) + (8-4) + (10-4)}{5} = 2 \end{aligned}$$

(b)

$$\begin{aligned} m'_2 &= \overline{(X - 4)^2} = \frac{\sum (X - 4)^2}{N} \\ &= \frac{(2-4)^2 + (3-4)^2 + (7-4)^2 + (8-4)^2 + (10-4)^2}{5} = \frac{66}{5} = 13.2 \end{aligned}$$

(c)

$$\begin{aligned} m'_3 &= \overline{(X - 4)^3} = \frac{\sum (X - 4)^3}{N} \\ &= \frac{(2-4)^3 + (3-4)^3 + (7-4)^3 + (8-4)^3 + (10-4)^3}{5} = \frac{298}{5} = 59.6 \end{aligned}$$

(d)

$$\begin{aligned} m'_4 &= \overline{(X - 4)^4} = \frac{\sum (X - 4)^4}{N} \\ &= \frac{(2-4)^4 + (3-4)^4 + (7-4)^4 + (8-4)^4 + (10-4)^4}{5} \\ &= \frac{1650}{5} = 330 \end{aligned}$$

5.4 根据习题 5.2 和 5.3 的结论验证下列关系:

(a) $m_2 = m'_2 - m_1'^2$, (b) $m_3 = m'_3 - 3m_1' m'_2 + 2m_1'^3$,

(c) $m_4 = m'_4 - 4m_1' m'_3 + 6m_1'^2 m'_2 - 3m_1'^4$.

解 根据习题 5.3, 有 $m_1' = 2$, $m_2' = 13.2$, $m_3' = 59.6$, $m_4' = 330$. 所以

$$(a) m_2 = m'_2 - m_1'^2 = 13.2 - 2^2 = 13.2 - 4 = 9.2$$

$$(b) m_3 = m'_3 - 3m'_1 m'_2 + 2m_1'^3 = 59.6 - 3 \times 2 \times 13.2 + 2 \times 2^3 = 59.6 - 79.2 + 16 = -3.6$$

$$(c) m_4 = m'_4 - 4m'_1 m'_3 + 6m_1'^2 m'_2 - 3m_1'^4 = 330 - 4 \times 2 \times 59.6 + 6 \times 2^2 \times 13.2 - 3 \times 2^4 = 122$$

这与习题 5.2 的结论是一致的.

5.5 证明 (a) $m_2 = m'_2 - m_1'^2$, (b) $m_3 = m'_3 - 3m'_1 m'_2 + 2m_1'^3$,

$$(c) m_4 = m'_4 - 4m'_1 m'_3 + 6m_1'^2 m'_2 - 3m_1'^4.$$

证明 如果 $d = X - A$, 那么 $X = A + d$, $\bar{X} = A + \bar{d}$, $X - \bar{X} = d - \bar{d}$. 因此

$$(a) m_2 = \overline{(X - \bar{X})^2} - \overline{(d - \bar{d})^2} = \overline{d^2 - 2d\bar{d} + \bar{d}^2} = \overline{d^2} - 2\bar{d}\overline{d} + \bar{d}^2 = \overline{d^2} - \bar{d}^2 = m'_2 - m_1'^2$$

$$(b) m_3 = \overline{(X - \bar{X})^3} - \overline{(d - \bar{d})^3} = \overline{d^3 - 3d\bar{d}^2 + 3\bar{d}d^2 - \bar{d}^3} = \overline{d^3} - 3\bar{d}\overline{d^2} + 3\bar{d}^2\overline{d} - \bar{d}^3 = m'_3 - 3m'_1 m'_2 + 2m_1'^3$$

$$(c) m_4 = \overline{(X - \bar{X})^4} - \overline{(d - \bar{d})^4} = \overline{d^4 - 4d\bar{d}^3 + 6d^2\bar{d}^2 - 4\bar{d}d^3 + \bar{d}^4} = \overline{d^4} - 4\bar{d}\overline{d^3} + 6\bar{d}^2\overline{d^2} - 4\bar{d}^3\overline{d} + \bar{d}^4 = m'_4 - 4m'_1 m'_3 + 6m_1'^2 m'_2 - 3m_1'^4$$

由这种思想, 我们也可得到 m_5, m_6 等的相似结果.

分类资料矩的计算

5.6 求习题 3.22 中身高分布的前四阶中心矩.

解 数据整理在表 5.1 中, 从中我们可以看出:

$$m'_1 = c \frac{\sum fu}{N} = 3 \times \frac{15}{100} = 0.45 \quad m'_3 = c^3 \frac{\sum fu^3}{N} = 3^3 \times \frac{33}{100} = 8.91$$

$$m'_2 = c^2 \frac{\sum fu^2}{N} = 3^2 \times \frac{97}{100} = 8.73 \quad m'_4 = c^4 \frac{\sum fu^4}{N} = 3^4 \times \frac{253}{100} = 204.93$$

因此

$$m_1 = 0$$

$$m_2 = m'_2 - m_1'^2 = 8.73 - 0.45^2 = 8.5275$$

$$m_3 = m'_3 - 3m'_1 m'_2 + 2m_1'^3 = 8.91 - 3 \times 0.45 \times 8.73 + 2 \times 0.45^3 = -2.6932$$

$$m_4 = m'_4 - 4m'_1 m'_3 + 6m_1'^2 m'_2 - 3m_1'^4 = 204.9 - 4 \times 0.45 \times 8.91 + 6 \times 0.45^2 \times 8.73 - 3 \times 0.45^4 = 199.3759$$

表 5.1

X	u	f	fu	fu ²	fu ³	fu ⁴
61	-2	5	-10	20	-40	80
64	-1	18	-18	18	-18	18
67	0	42	0	0	0	0
70	1	27	27	27	27	27
73	2	8	16	32	64	128
		$N = \sum f = 10$	$\sum fu = 15$	$\sum fu^2 = 97$	$\sum fu^3 = 33$	$\sum fu^4 = 253$

5.7 求习题 4.19 表 4.7 中分布的 (a) m'_1 , (b) m'_2 , (c) m'_3 , (d) m'_4 , (e) m_1 , (f) m_2 , (g) m_3 , (h) m_4 , (i) \bar{X} , (j) s , (k) $\overline{X^2}$, (l) $\overline{X^3}$.

解 数据整理在表 5.2 中.

表 5.2

X	u	f	fu	fu^2	fu^3	fu^4
70	-6	4	-24	144	-864	5184
74	-5	9	-45	225	-1125	5625
78	-4	16	-64	256	-1024	4096
82	-3	28	-84	252	-756	2268
86	-2	45	-90	180	-360	720
90	-1	66	-66	66	-66	66
$A \rightarrow 94$	0	85	0	0	0	0
98	1	72	72	72	72	72
102	2	54	108	216	432	864
106	3	38	114	342	1026	3078
110	4	27	108	432	1728	6912
114	5	18	90	450	2250	11250
118	6	11	66	396	2376	14256
122	7	5	34	245	1715	12005
126	8	2	16	128	1024	8192
		$N = \sum f = 480$	$\sum fu = 236$	$\sum fu^2 = 3404$	$\sum fu^3 = 6428$	$\sum fu^4 = 74\ 588$

$$(a) m'_1 = c \frac{\sum fu}{N} = 4 \times \frac{236}{480} = 1.9667$$

$$(b) m'_2 = c^2 \frac{\sum fu^2}{N} = 4^2 \times \frac{3404}{480} = 113.4667$$

$$(c) m'_3 = c^3 \frac{\sum fu^3}{N} = 4^3 \times \frac{6428}{480} = 857.0667$$

$$(d) m'_4 = c^4 \frac{\sum fu^4}{N} = 4^4 \times \frac{74\ 588}{480} = 39\ 780.2667$$

$$(e) m_1 = 0$$

$$(f) m_2 = m'_2 - m_1'^2 = 113.4667 - 1.9667^2 = 109.5988$$

$$(g) m_3 = m'_3 - 3m'_1 m'_2 + 2m_1'^3$$

$$= 857.0667 - 3 \times 1.9667 \times 113.4667 + 2 \times 1.9667^3 = 202.8158$$

$$(h) m_4 = m'_4 - 4m'_1 m'_3 + 6m_1'^2 m'_2 - 3m_1'^4 = 35\ 627.2853$$

$$(i) \bar{X} = \overline{A + d} = A + m'_1 = A + c \frac{\sum fu}{N} = 94 + 1.9667 = 95.97$$

$$(j) s = \sqrt{m_2} = \sqrt{109.5988} = 10.47$$

$$(k) \overline{X^2} = \overline{(A + d)^2} = \overline{A^2 + 2Ad + d^2} = A^2 + 2A\bar{d} + \bar{d}^2 = A^2 + 2Am'_1 + m'_2$$

$$= 94^2 + 2 \times 94 \times 1.9667 + 113.4667 = 9319.2063 \text{ 或(保留四个有效数字) } 9319$$

$$(l) \overline{X^3} = \overline{(A + d)^3} = \overline{A^3 + 3A^2d + 3Ad^2 + d^3} = A^3 + 3A^2\bar{d} + 3A\bar{d}^2 + \bar{d}^3$$

$$= A^3 + 3A^2m'_1 + 3Am'_2 + m'_3 = 915\ 571.9597 \text{ 或(保留四个有效数字) } 915\ 600.$$

Charlier 检验

5.8 解释在习题 5.7 计算中 Charlier 检验的作用.

解 为了应用 Charlier 检验, 表 5.3 由表 5.2 而得(为了方便, 第二列照写).

在以下计算中, 第一式取自于表 5.3, 第二式取自于表 5.2. 每组计算就给出了所要求的检验.

表 5.3

$u+1$	f	$f(u+1)$	$f(u+1)^2$	$f(u+1)^3$	$f(u+1)^4$
-5	4	-20	100	-500	2500
-4	9	-36	144	-576	2304
-3	16	-48	144	-432	1296
-2	28	-56	112	-224	448
-1	45	-45	45	-45	45
0	66	0	0	0	0
1	85	85	85	85	85
2	72	144	288	576	1152
3	54	162	486	1458	4374
4	38	152	608	2432	9728
5	27	135	675	3375	16875
6	18	108	648	3888	23328
7	11	77	539	3773	26411
8	5	40	320	2560	20480
9	2	18	162	1458	13122
$N = \sum f = 480$		$\sum f(u+1) = 716$	$\sum f(u+1)^2 = 4356$	$\sum f(u+1)^3 = 17\,828$	$\sum f(u+1)^4 = 122\,148$

$$\sum f(u+1) = 716$$

$$\sum fu + N = 236 + 480 = 716$$

$$\sum f(u+1)^2 = 4356$$

$$\sum fu^2 + 2 \sum fu + N = 3404 + 2 \times 236 + 480 = 4356$$

$$\sum f(u+1)^3 = 17\,828$$

$$\sum fu^3 + 3 \sum fu^2 + 3 \sum fu + N = 6428 + 3 \times 3404 + 3 \times 236 + 480 = 17\,828$$

$$\sum f(u+1)^4 = 122\,148$$

$$\sum fu^4 + 4 \sum fu^3 + 6 \sum fu^2 + 4 \sum fu + N = 74\,588 + 4 \times 6428 + 6 \times 3404 + 4 \times 236 + 480 = 122\,148$$

Sheppard 矩修正

5.9 应用 Sheppard 修正求(a)习题 5.6, (b)习题 5.7 中数据的中心矩.

解 (a)修正的 $m_2 = m_2 - c^2/12 = 8.5275 - 3^2/12 = 7.7775$

$$\begin{aligned} \text{修正的 } m_4 &= m_4 - \frac{1}{2}c^2m_2 + \frac{7}{240}c^4 \\ &= 199.3759 - \frac{1}{2} \times 3^2 \times 8.5275 + \frac{7}{240} \times 3^4 \\ &= 163.3646 \end{aligned}$$

m_1, m_2 无需修正.

(b)修正的 $m_2 = m_2 - c^2/12 = 109.5988 - 4^2/12 = 108.2655$

$$\text{修正的 } m_4 = m_4 - \frac{1}{2}c^2m_2 + \frac{7}{240}c^4$$

$$\begin{aligned}
&= 35627.2853 - \frac{1}{2} \times 4^2 \times 109.5988 + \frac{7}{240} \times 4^4 \\
&= 34\,757.9616
\end{aligned}$$

偏度

- 5.10 求 P&R 公司 65 个员工工资分布的 Pearson(a) 第一, (b) 第二偏度系数(见习题 3.44 和 4.18).

解 均值 = 279.76 美元, 中位数 = 279.06 美元, 众数 = 277.50 美元, 标准差 $s = 15.60$ 美元. 因此,

$$(a) \text{ 第一偏度系数} = \frac{\text{均值} - \text{众数}}{s} = \frac{279.76 - 277.50}{15.60} = 0.1448 \text{ 或 } 0.14$$

$$(b) \text{ 第二偏度系数} = \frac{3(\text{均值} - \text{中位数})}{s} = \frac{3(279.76 - 279.06)}{15.60} = 0.1346 \text{ 或 } 0.13$$

如果应用修正的标准差(见习题 4.21(b)), 这些系数分别变为

$$(a) \frac{\text{均值} - \text{众数}}{\text{修正的 } s} = \frac{279.76 - 277.50}{15.33} = 0.1474 \text{ 或 } 0.15$$

$$(b) \frac{3(\text{均值} - \text{中位数})}{\text{修正的 } s} = \frac{3(279.76 - 279.06)}{15.33} = 0.1370 \text{ 或 } 0.14$$

由于系数是正的, 分布有正偏度(即向右偏).

- 5.11 求习题 5.10 中分布的(a)四分位, (b)百分位偏度系数(见习题 3.44).

解 $Q_1 = 268.25$ 美元, $Q_2 = P_{50} = 279.06$ 美元, $Q_3 = 290.75$ 美元, $P_{10} = D_1 = 258.12$ 美元, $P_{90} = D_9 = 301.00$ 美元.

$$(a) \text{ 四分位偏度系数} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} = \frac{290.75 - 2 \times 279.06 + 268.25}{290.75 - 268.25} = 0.0391$$

$$(b) \text{ 百分位偏度系数} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} = \frac{301.00 - 2 \times 279.06 + 258.12}{301.00 - 258.12} = 0.0233$$

- 5.12 求(a)XYZ 大学学生身高分布(见习题 5.6), (b)小学生 IQ(见习题 5.7)的矩偏度系数 a_3 .

解 (a) $m_2 = s^2 = 8.5275$, $m_3 = -2.6932$. 因此

$$a_3 = \frac{m_3}{s^3} = \frac{m_3}{(\sqrt{m_2})^3} = \frac{-2.6932}{(\sqrt{8.5275})^3} = -0.1081 \text{ 或 } -0.11$$

如果应用 Sheppard 修正(见习题 5.9(a)), 那么

$$\text{修正的 } a_3 = \frac{m_3}{(\sqrt{\text{修正的 } m_2})^3} = \frac{-2.6932}{(\sqrt{7.7775})^3} = -0.1242 \text{ 或 } -0.12$$

$$(b) a_3 = \frac{m_3}{s^3} = \frac{m_3}{(\sqrt{m_2})^3} = \frac{202.8158}{(\sqrt{109.5988})^3} = 0.1768 \text{ 或 } 0.18$$

如果应用 Sheppard 修正(见习题 5.9(b)), 那么

$$\text{修正的 } a_3 = \frac{m_3}{(\sqrt{\text{修正的 } m_2})^3} = \frac{202.8158}{(\sqrt{108.2655})^3} = 0.1800 \text{ 或 } 0.18$$

注意这里的两个分布都是微斜的, (a) 中分布向左斜(负偏度), (b) 中分布向右斜(正偏度). (b) 中分布比(a)中分布倾斜些, 即(a)中分布比(b)中分布对称些, 从(b)的偏度系数绝对值比(a)的偏度系数绝对值大这一事实中也可看出.

峰度

- 5.13 求(a)习题 5.6, (b)习题 5.7 中数据的矩峰度系数 a_4 .

$$\text{解 (a) } a_4 = \frac{m_4}{s^4} = \frac{m_4}{m_2^2} = \frac{199.3759}{8.5275^2} = 2.7418 \text{ 或 } 2.74$$

如果应用 Sheppard 修正(见习题 5.9(a)), 那么

$$\text{修正的 } a_4 = \frac{\text{修正的 } m_4}{(\text{修正的 } m_2)^2} = \frac{163.36346}{(7.7775)^2} = 2.7007 \text{ 或 } 2.70$$

$$(b) \quad a_4 = \frac{m_4}{s^4} = \frac{m_4}{m_2^2} = \frac{35\,627.2853}{109.5988^2} = 2.9660 \text{ 或 } 2.97$$

如果应用 Sheppard 修正(见习题 5.9(b)), 那么

$$\text{修正的 } a_4 = \frac{\text{修正的 } m_4}{(\text{修正的 } m_2)^2} = \frac{34\,757.9616}{108.2655^2} = 2.9653 \text{ 或 } 2.97$$

由于正态分布 $a_4 = 3$, 因此(a)、(b)两个分布相对于正态分布都是**扁峰**的(即没有正态分布那么陡峭).

考虑到陡峭程度, (b)近似正态分布的程度比(a)好. 然而根据习题 5.12, (a)中分布比(b)中对称些, 因此考虑到对称性, (a)近似正态分布的程度比(b)好.

- 5.14 (a) 计算习题 5.11 中分布的百分位峰度系数 $\kappa = Q/(P_{90} - P_{10})$;
(b) 讨论它与正态分布的近似程度.

解 (a) $Q = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2} \times (290.75 - 268.25) = 11.25$ 美元, $P_{90} - P_{10} = 301.00 - 258.12 = 42.88$ 美元, 因此 $\kappa = Q/(P_{90} - P_{10}) = 0.262$.

(b) 由于正态分布的 κ 是 0.263, 因此给定分布是**常峰态**的(即与正态分布差不多尖), 分布的峰度应与正态分布的大致一样. 如果从峰度来考虑, 我们相信它与正态分布是近似的.

补充习题

矩

- 5.15 求数集 4, 7, 5, 9, 8, 3, 6 的(a)一阶矩, (b)二阶矩, (c)三阶矩, (d)四阶矩.
5.16 求习题 5.15 中数据的(a)一阶中心矩, (b)二阶中心矩, (c)三阶中心矩, (d)四阶中心矩.
5.17 求习题 5.15 中数据的关于原点 7 的(a)一阶矩, (b)二阶矩, (c)三阶矩, (d)四阶矩.
5.18 根据习题 5.16 和 5.17 的结论验证下列关系:
(a) $m_2 = m'_2 - m_1^2$, (b) $m_3 = m'_3 - 3m'_1m'_2 + 2m_1^3$, (c) $m_4 = m'_4 - 4m'_1m'_3 + 6m_1^2m'_2 - 3m_1^4$.
5.19 求等差数列 2, 5, 8, 11, 14, 17 的前四阶中心矩.
5.20 证明(a) $m'_2 = m_2 + h^2$, (b) $m'_3 = m_3 + 3hm_2 + h^3$,
(c) $m'_4 = m_4 + 4hm_3 + 6h^2m_2 + h^4$, 其中, $h = m'_1$.
5.21 如果关于原点 2 的一阶矩等于 5, 那么均值是多少?
5.22 如果关于原点 3 的前四阶矩分别为 -2, 10, -25, 50, 求相应的(a)中心矩, (b)关于原点 5 的矩, (c)原点矩(即关于原点 0 的矩).
5.23 求 0, 0, 0, 1, 1, 1, 1 和 1 的前四阶中心矩.
5.24 (a) 证明: $m_5 = m'_5 - 5m'_1m'_4 + 10m_1^2m'_3 - 10m_1^3m'_2 + 4m_1^5$.
(b) 写出一个关于 m_6 的类似的公式.
5.25 有 N 个取 1 或 0 的数, 其中取“1”与取“0”的个数之比为 $p:q$ ($q = 1 - p$). 求这组数的(a) m_1 , (b) m_2 , (c) m_3 , (d) m_4 , 并同习题 5.23 相比较.
5.26 证明等差数列 $a, a + d, a + 2d, \dots, a + (n - 1)d$ 的前四阶中心矩为: $m_1 = 0$, $m_2 = \frac{1}{12}(n^2 - 1)d^2$, $m_3 = 0$,
 $m_4 = \frac{1}{240}(n^2 - 1)(3n^2 - 7)d^4$. 与习题 5.19 作比较(见习题 4.69). (提示: $1^4 + 2^4 + 3^4 + \dots + (n - 1)^4 = \frac{1}{30}n(n - 1)(2n - 1)(3n^2 - 3n - 1)$).

分类资料的矩

- 5.27 根据表 5.4 的分布求相应的前四阶中心矩.

表 5.4

X	f
12	1
14	4
16	6

续表

X	f
18	10
20	7
22	2
总计	30

- 5.28 解释在习题 5.27 计算中 Charlier 检验的作用.
- 5.29 用 Sheppard 修正习题 5.27 得到的矩.
- 5.30 (a)不用 Sheppard 修正, (b)用 Sheppard 修正计算习题 3.59 中分布的前四阶中心矩.
- 5.31 根据习题 3.62 中分布计算 (a) m_1 , (b) m_2 , (c) m_3 , (d) m_4 , (e) \bar{X} , (f) s , (g) $\overline{X^2}$, (h) $\overline{X^3}$, (i) $\overline{X^4}$, (j) $\overline{(X+1)^3}$.

偏度

- 5.32 (a)不用 Sheppard 修正, (b)用 Sheppard 修正计算习题 5.27 中分布的矩偏度系数 a_3 .
- 5.33 计算习题 3.59(见习题 5.30)中分布的矩偏度系数 a_3 .
- 5.34 两个分布的二阶中心矩分别为 9 和 16, 三阶中心矩分别为 -8.1 和 -12.8. 请问哪个分布向左斜得更多?
- 5.35 求习题 3.59 的 Pearson(a)第一, (b)第二偏度系数, 并解释出现的差异.
- 5.36 求习题 3.59 的 (a)四分位和 (b)百分位偏度系数. 与习题 5.35 的结果作比较, 并作出解释.
- 5.37 表 5.5 给出变量 X 的三种不同分布. 这三种分布的频数分别为 f_1, f_2 和 f_3 . 对三种分布求 Pearson 第一, 第二偏度系数. 在计算系数时, 可应用修正标准差.

表 5.5

X	f_1	f_2	f_3
0	10	1	1
1	5	2	2
2	2	14	2
3	2	2	5
4	1	1	10

峰度

- 5.38 (a)不用 Sheppard 修正, (b)用 Sheppard 修正计算习题 5.27(a)中分布的矩峰度系数.
- 5.39 (a)不用 Sheppard 修正, (b)用 Sheppard 修正计算习题 3.59(a)中分布的矩峰度系数.
- 5.40 习题 5.34 中两个分布的四阶中心矩分别为 230 和 780. 请问从 (a)峰度, (b)偏度上看哪个分布更近似于正态分布?
- 5.41 习题 5.40 中的分布哪一个是 (a)尖峰的, (b)常峰态的, (c)扁峰的?
- 5.42 一个对称分布的标准差是 5. 如果要使得分布是 (a)尖峰的, (b)常峰态的, (c)扁峰的, 那么四阶中心矩的值应是多少?
- 5.43 (a)计算习题 3.59 中分布的百分位峰度系数 κ .
- (b)把你的结果与正态分布的理论值 0.263 作比较, 并解释原因.
- (c)你如何把这个结论与习题 5.39 的结论统一起来?

第六章 初等概率论

概率的定义

概率的古典定义

假设事件 E 在等可能的 n 种方式中可以以 h 种方式发生, 则事件发生(成功)的概率表示为

$$p = P(E) = \frac{h}{n}$$

此事件不发生(失败)的概率表示为

$$q = P(\text{非 } E) = \frac{n-h}{n} = 1 - \frac{h}{n} = 1 - p = 1 - P(E)$$

则 $p + q = 1$ 或 $P(E) + P(\text{非 } E) = 1$. 事件“非 E ”有时可记为 \bar{E} .

例 1 设 E 为事件“掷一次骰子出现 3 或 4”. 骰子共有六种下落方式, 分别出现 1, 2, 3, 4, 5 或 6, 如果骰子是均匀的(即没有负重), 我们可假设六种下落方式是等可能的. 因为 E 可以以其中两种方式发生, 所以 $p = P(E) = \frac{2}{6} = \frac{1}{3}$. 3 或 4 不出现(即 1, 2, 5 或 6 出现)的概率为 $q = P(\bar{E}) = 1 - \frac{1}{3} = \frac{2}{3}$.

注意, 事件发生的概率介于 0 和 1 之间. 如果事件必然不发生, 它的概率为 0. 如果必然发生(即发生是肯定的), 它的概率为 1.

如果一个事件要发生的概率为 p , 则赞同它发生的胜败之比为 $p:q$ (读作“ p 比 q ”); 否定它发生的胜败之比为 $q:p$. 这样掷一个均匀的骰子, 赞同 3 或 4 不出现的胜败之比为 $q:p = \frac{2}{3}:\frac{1}{3} = 2:1$ (即 2 比 1).

概率的统计定义

概率的古典定义有所缺陷, “等可能”这一词模糊不清. 事实上, 这一词看上去和“等概率”是同义的, 那么我们实质上是用概率来定义它自己, 这就成了循环定义. 为此, 有人提出要给概率下一个统计学上的定义. 据此, 当观测次数很大时, 事件的估计概率或**经验概率**被理解为事件发生的**频率**. 统计定义的概率即为观测次数无限增大时频率的极限.

例 2 如果抛掷硬币 1000 次有 529 次出现正面, 则出现正面的相对频率为 $529/1000 = 0.529$. 再抛掷 1000 次有 493 次出现正面, 则抛掷 2000 次出现正面的频率为 $(529 + 493)/2000 = 0.511$. 根据统计上的定义, 这样继续做下去, 频率最终会和一个数字越来越接近, 它就是抛掷一次硬币出现正面的概率. 从上述结果可见, 这个数字应该在 0.5 与某个有效数字之间. 要想获得更有意义的结论, 还要做更多的观测.

虽然统计上的定义在实际应用中是可以的, 但从数学的观念上来说仍存在问题, 因为实际的极限值有可能根本不存在. 为此, 现代概率论已在向**公理化**的方向发展; 也就是说, 并没有给概率在理论上加以定义, 就如同几何学中**点**和**线**没有定义一样.

条件概率, 独立和不独立事件

假设有 E_1, E_2 两个事件, 在已知 E_1 发生的条件下 E_2 发生的概率记为 $P(E_2|E_1)$, 这个概率称作是在 E_1 发生的条件下, E_2 发生的**条件概率**.

如果 E_1 是否发生并不影响 E_2 发生的概率, 则 $P(E_2|E_1) = P(E_2)$, 此时称 E_1 和 E_2 是

独立事件;否则,它们是不独立事件.

如果我们用 E_1E_2 表示事件“ E_1 和 E_2 同时发生”,有时称**复合事件**,则

$$P(E_1E_2) = P(E_1)P(E_2 | E_1) \quad (1)$$

特别地,对于独立事件,

$$P(E_1E_2) = P(E_1)P(E_2) \quad (2)$$

对于三个事件 E_1, E_2 和 E_3 ,我们有

$$P(E_1E_2E_3) = P(E_1)P(E_2 | E_1)P(E_3 | E_1E_2) \quad (3)$$

也就是说, E_1, E_2 和 E_3 同时发生的概率等于

(E_1 发生的概率) \times (E_1 发生条件下 E_2 发生的条件概率) \times (E_1, E_2 同时发生条件下 E_3 发生的条件概率)

特别地,对于独立事件,

$$P(E_1E_2E_3) = P(E_1)P(E_2)P(E_3) \quad (4)$$

一般说来,如果 $E_1, E_2, E_3, \dots, E_n$ 是 n 个独立事件,它们发生的概率分别为 $p_1, p_2, p_3, \dots, p_n$,那么 E_1, E_2, \dots, E_n 同时发生的概率是 $p_1p_2p_3 \cdots p_n$.

例 3 多次抛掷一枚硬币, E_1 和 E_2 分别表示事件“第五次出现正面”和“第六次出现正面”.则 E_1 和 E_2 是独立事件,且第五次和第六次都出现正面(假设硬币是均匀的)的概率为

$$P(E_1E_2) = P(E_1)P(E_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

例 4 如果 A 将存活 20 年的概率为 0.7, B 将存活 20 年的概率为 0.5,那么它们都将存活 20 年的概率为 $0.7 \times 0.5 = 0.35$.

例 5 假设盒子里有 3 个白球和 2 个黑球,从中无放回地取球.设 E_1 为事件“第一次取出的是黑球”, E_2 为事件“第二次取出的是黑球”. E_1 和 E_2 是不独立事件.

第一次取出的是黑球的概率为 $P(E_1) = \frac{2}{3+2} = \frac{2}{5}$. 在第一次取出黑球的条件下第二次取出黑球的概率为 $P(E_2 | E_1) = \frac{1}{3+1} = \frac{1}{4}$,那么两次取出的都是黑球的概率为

$$P(E_1E_2) = P(E_1)P(E_2 | E_1) = \frac{2}{5} \cdot \frac{1}{4} = \frac{1}{10}$$

互不相容事件

如果两个或多个事件中的任意两个事件都不能同时发生,那么称它们是**互不相容的**. 因此若 E_1 和 E_2 是互不相容事件,则 $P(E_1E_2) = 0$.

如果用 $E_1 \cup E_2$ 表示事件“ E_1 和 E_2 中至少有一个发生”,则

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1E_2) \quad (5)$$

特别地,对于互不相容事件,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) \quad (6)$$

加以推广可知,如果 E_1, E_2, \dots, E_n 是 n 个互不相容的事件,它们发生的概率分别为 p_1, p_2, \dots, p_n ,则 E_1, E_2, \dots, E_n 中至少有一个发生的概率为 $p_1 + p_2 + \dots + p_n$.

结果(5)也可推广到三个或更多个事件(见习题 6.38).

例 6 从一副纸牌中抽牌一次,如果 E_1 为事件“抽到 A”, E_2 为事件“抽到 K”,则 $P(E_1) = \frac{4}{52} = \frac{1}{13}$, $P(E_2) = \frac{4}{52} = \frac{1}{13}$. 因为不可能同时抽到 A 和 K,所以它们是互不相容的,则抽到 A 或 K 的概率为

$$P(E_1 + E_2) = P(E_1) + P(E_2) = \frac{1}{13} + \frac{1}{13} = \frac{2}{13}$$

例 7 从一副纸牌中抽牌一次, 如果 E_1 为事件“抽到 A”, E_2 为事件“抽到黑桃”. 因为有可能抽到黑桃 A, 所以 E_1 和 E_2 不是互不相容的. 那么抽到 A 或黑桃的概率为

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 E_2) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

概率分布

离散型

如果变量 X 取到一组离散值 X_1, X_2, \dots, X_k 的概率分别为 p_1, p_2, \dots, p_k , 其中 $p_1 + p_2 + \dots + p_k = 1$, 我们则称对 X 定义了一个**离散型概率分布**. 当 $X = X_1, X_2, \dots, X_k$ 时函数 $p(X)$ 分别取值 p_1, p_2, \dots, p_k , 则称 $p(X)$ 为 X 的**概率函数**或**频率函数**. 由于 X 可以以给定概率取到确定的数值, 它常常被称为**离散型随机变量**. 随机变量也称为**机会变量**或**随机变数**.

例 8 投掷一对骰子, 用 X 表示所得数字之和. X 的概率分布见表 6.1. 例如, 所得数字之和为 5 的概率为 $\frac{4}{36} = \frac{1}{9}$; 因此, 投掷一对骰子 900 次我们预期会有 100 次得到数字之和为 5.

表 6.1

X	2	3	4	5	6	7	8	9	10	11	12
$p(X)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

注意, 这有点类似于将频率分布中的频率换为概率, 所以我们可以认为概率分布就是当观测次数很多时频率分布的理论上的或理想中的极限形式. 因此, 我们可以认为概率分布是**总体**的分布, 而频率分布是总体中**样本**的分布.

就像频率分布一样, 概率分布也可用 $p(X)$ 对 X 的图形表示出来(见习题 6.11).

类似于累积频率分布, 我们也可以通过累积概率得到**累积概率分布**. 与此分布相关的函数有时称为**分布函数**.

连续型

上述思想也可推广到变量 X 取连续值的情形. 从理论上或极限上来看, 频率多边形变成了一条连续的曲线(如图 6-1), 用方程 $Y = p(X)$ 来表示. 曲线下方和 X 轴所围的面积为 1, 曲线下方和 X 轴上方在直线 $X = a$ 及 $X = b$ 之间区域面积(图 6-1 中阴影部分)表示 X 介于 a 和 b 之间的概率, 可记为 $P(a < X < b)$.

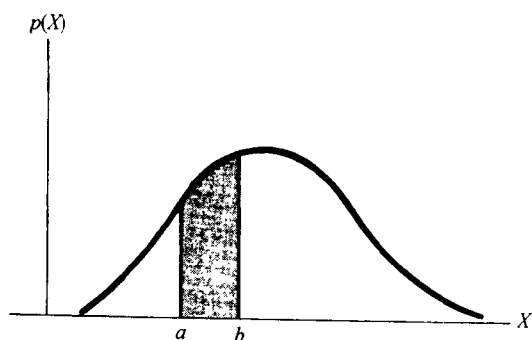


图 6-1

我们称 $p(X)$ 为**概率密度函数**, 简称为**密度函数**. 当密度函数给定后, 我们就可以说对 X 定义了一个**连续型概率分布**. 变量 X 被称为**连续型随机变量**.

和离散型情形一样,我们也可定义累积概率分布以及相关的分布函数.

数学期望

如果一个人总共获得 S 美元的的概率为 p , 则他的**数学期望**(或简称**期望**)定义为 pS .

例 9 一个人获得 10 美元奖金的概率为 $\frac{1}{5}$, 那么他的期望是 $\frac{1}{5} \times 10$ 美元 = 2 美元.

可以很容易地将期望的概念加以扩展. 如果 X 表示一离散型随机变量, 它分别以概率 p_1, p_2, \dots, p_k 取到值 X_1, X_2, \dots, X_k , 其中 $p_1 + p_2 + \dots + p_k = 1$, 则 X 的数学期望(简称为 X 的期望或均值)用 $E(X)$ 表示, 定义如下

$$E(X) = p_1 X_1 + p_2 X_2 + \dots + p_k X_k = \sum_{j=1}^k p_j X_j = \sum pX \quad (7)$$

如果用频率 f_j/N 代替期望中的概率 p_j , 其中 $N = \sum f_j$, 则期望变为 $(\sum fX)/N$, 这正是 X_1, X_2, \dots, X_k 分别以这些频率出现的一个容量为 N 的样本的均值 \bar{X} . 当 N 越来越大时, 频率 f_j/N 越来越接近概率 p_j . 因此我们可以把 $E(X)$ 理解为提供样本来源的总体的均值. 如果样本的均值记为 m , 则可用相应的希腊字母 μ 表示总体的均值.

也可定义连续型随机变量的期望, 但定义中要用到微积分.

总体均值和方差与样本均值和方差的关系

如果我们从总体中随机抽取一个容量为 N 的样本(即我们假定抽到各样品是等概率地), 则**样本均值 m 的期望值是总体均值 μ** .

但是这并不表明样本的任何量的期望值都等于总体的相应量. 比如说, 我们前面定义的样本方差的期望值就不等于总体的方差, 而是方差乘以 $(N-1)/N$. 因此有些统计学家就定义样本方差是普通方差乘以 $N/(N-1)$.

组合分析

为了求得复杂事件发生的概率, 常常要列举不同的情况, 而这些都是困难而烦琐的. 要简化这些工作, 就要用到组合分析中的一些基本原理.

基本原理

如果一个事件可以以 n_1 种方式中的任一种方式发生, 并且该事件发生时引起另一事件以 n_2 种方式中的任一种方式发生, 则两个事件可以按某种特定顺序都发生的方式有 $n_1 n_2$ 种.

例 10 如果有 3 个人竞选省长, 5 个人竞选市长, 那么产生这两个职位的方法共有 $3 \times 5 = 15$ 种.

n 的阶乘

n 的阶乘记为 $n!$, 定义为

$$n! = n(n-1)(n-2)\cdots 1 \quad (8)$$

则 $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$, $4! \times 3! = (4 \times 3 \times 2 \times 1) \times (3 \times 2 \times 1) = 144$. 为方便起见, 定义 $0! = 1$.

排列

从 n 个不同物体中一次取出 r 个物体按照一定的顺序排成一列称为从 n 个物体中取出 r 个物体的**排列**, 简称为 n 中取 r 的排列. n 中取 r 的排列数记为 ${}_nP_r$, $P(n, r)$ 或 $P_{n,r}$, 定义为

$${}_nP_r = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!} \quad (9)$$

特别地, n 中取 n 的排列数为

$${}_nP_n = n(n-1)(n-2)\cdots 1 = n!$$

例 11 从字母 a, b 和 c 中一次取出两个的排列数是 ${}_3P_2 = 3 \times 2 = 6$, 它们是 ab, ba, ac, ca, bc 和 cb .

如果 n 个物体中分别有 n_1, n_2, \dots, n_k 个是相同的, 则 n 个物体的排列数是

$$\frac{n!}{n_1! n_2! \cdots n_k!} \quad (10)$$

其中 $n = n_1 + n_2 + \cdots + n_k$.

例 12 在单词 statistics 中有 3 个 s , 3 个 t , 1 个 a , 2 个 i 和 1 个 c , 则这个单词中的字母的排列数是

$$\frac{10!}{3! 3! 1! 2! 1!} = 50\,400$$

组合

从 n 个不同的物体中一次取出 r 个物体组成一组而不考虑它们的顺序如何, 称为从 n 个物体中取 r 个物体的组合, 简称为 n 中取 r 的组合. n 中取 r 的组合数记为 $\binom{n}{r}$, 定义为

$$\binom{n}{r} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!} \quad (11)$$

例 13 从字母 a, b 和 c 中一次取出两个的组合数是

$$\binom{3}{2} = \frac{3 \times 2}{2!} = 3$$

它们是 ab, ac 和 bc . 注意到 ab 和 ba 是相同的组合, 但却是不同的排列.

$n!$ 的 Stirling 逼近

当 n 很大时, 直接计算 $n!$ 是不切实际的. 这时, 要用到 James Stirling 的逼近公式:

$$n! \approx \sqrt{2\pi n} n^n e^{-n} \quad (12)$$

其中 $e = 2.71828\cdots$ 是自然对数的底 (见习题 6.31).

概率和集合论的关系

在现代概率论中, 我们把一次试验的所有可能结局 (或结果) 都看成样本空间 S (可以是一维, 二维, 三维……) 中的点. 如果 S 中只包含有限个点, 那么对于每一个点可用一个非负数与之相对应, 使得与所有点对应的数之和为 1, 这些数就称为 **概率**. 事件是 S 中点的集合 (或集), 如图 6-2 中 E_1 或 E_2 , 这种表示集合的图叫做 **欧拉图** 或 **文氏图**.

事件 $E_1 \cup E_2$ 是一个点集, 其中的点必至少居于 E_1, E_2 中的一个. 而事件 $E_1 E_2$ 是由 E_1 和 E_2 的公共部分构成的点集. 一个事件如 E_1 的概率就是集合 E_1 中所有点的概率之和. 类似地, $E_1 \cup E_2$ 的概率 $P(E_1 \cup E_2)$ 就是集合 $E_1 \cup E_2$ 中所有点的概率之和. 如果 E_1 和 E_2 没有公共点 (即事件是互不相容的), 则 $P(E_1 \cup E_2) = P(E_1) + P(E_2)$. 如果它们有公共点, 则 $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 E_2)$.

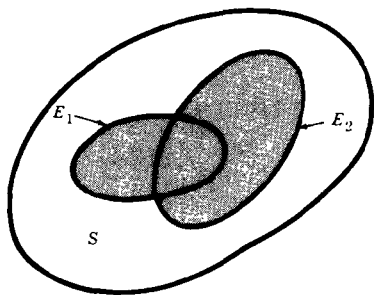


图 6-2

集合 $E_1 \cup E_2$ 称为两个集合的 **并**. 集合 $E_1 E_2$ 称为两个集合的 **交**, 有时也记为 $E_1 \cap E_2$. 推广到多于两个集合的情况, 我们可用记号 $E_1 \cap E_2 \cap E_3$ 代替

$$E_1 E_2 E_3.$$

符号 \emptyset 用于表示不含任何点的集合,称为**空集**.这种集合所对应的事件发生的概率为0(即 $P(\emptyset)=0$).如果 E_1 和 E_2 没有公共点,则写作 $E_1 E_2 = \emptyset$,也就是说它们对应的事件是互不相容的,因此 $P(E_1 E_2)=0$.

应用这些现代的观念,随机变量就是定义在样本空间的每一个点上的函数.例如,习题6.37中的随机变量是每个点的坐标之和.

当 S 中有无限个点时,可以用积分的概念将上述理论加以扩展.

习题及解答

概率的基本法则

6.1 求下列每个事件发生的概率 p 或其估计量:

- (a) 抛掷均匀骰子一次,出现奇数.
- (b) 抛掷均匀硬币两次,至少一次出现正面.
- (c) 从一副洗好的52张牌中抽取一张,抽到A或方块10或黑桃2.
- (d) 抛掷一对均匀的骰子一次,出现的数字和为7.
- (e) 如果抛掷硬币100次有56次正面向上,而下一次出现反面.

解 (a) 在六种等可能的情形中,该事件包括三种情形(1,3或5出现),则 $p = \frac{3}{6} = \frac{1}{2}$.

(b) 用 H 记“正面”, T 记“反面”,则抛掷两次会等可能地出现四种情形: HH, HT, TH 和 TT .该事件包括前三种情形,则 $p = \frac{3}{4}$.

(c) 在等可能的52种情形中,该事件可以以六种方式发生(黑桃A,红桃A,梅花A,方块A,方块10,黑桃2),则 $p = \frac{6}{52} = \frac{3}{26}$.

(d) 可以把一个骰子的六个面和另一个骰子的六个面结合起来看,因此能列举出 $6 \times 6 = 36$ 种等可能出现的情形,可记为 $(1,1), (2,1), (3,1), \dots, (6,6)$.其中有六种情形的数字和为7,分别为 $(1,6), (2,5), (3,4), (4,3), (5,2)$ 和 $(6,1)$ (见习题6.37(a)),则 $p = \frac{6}{36} = \frac{1}{6}$.

(e) 因为抛掷100次有 $100 - 56 = 44$ 次反面向上,所以下一次出现反面的估计(或经验)概率是出现反面的频率 $44/100 = 0.44$.

6.2 一次试验包括抛掷一枚硬币和一个骰子.用 E_1 表示事件“抛掷硬币时正面向上”,用 E_2 表示事件“抛掷骰子时3或6出现”,用语言叙述下列记号的意义:

- (a) \bar{E}_1 (b) \bar{E}_2 (c) $E_1 E_2$
- (d) $P(E_1 \bar{E}_2)$ (e) $P(E_1 | E_2)$ (f) $P(\bar{E}_1 \cup \bar{E}_2)$

解 (a) 不论骰子如何,抛掷硬币时反面出现;

(b) 不论硬币如何,抛掷骰子时1,2,4或5出现;

(c) 抛掷硬币时正面向上且抛掷骰子时3或6出现;

(d) 硬币正面向上且骰子出现1,2,4或5出现的概率;

(e) 在已知骰子出现3或6时硬币正面向上的概率;

(f) 硬币反面向上或骰子出现1,2,4或5的概率.

6.3 从一个装有6个红球,4个白球和5个蓝球的盒子里随机地抽取一个球.求取到下列球的概率:(a)红球,(b)白球,(c)蓝球,(d)不是红球,(e)红球或白球.

解 用 R, W 和 B 分别表示事件“抽到红球”,“抽到白球”和“抽到蓝球”.则

$$(a) P(R) = \frac{\text{抽到一个红球的可能方式数}}{\text{抽取一个球的所有可能方式数}} = \frac{6}{6+4+5} = \frac{6}{15} = \frac{2}{5}$$

$$(b) P(W) = \frac{4}{6+4+5} = \frac{4}{15}$$

$$(c) P(B) = \frac{5}{6+4+5} = \frac{5}{15} = \frac{1}{3}$$

$$(d) \text{ 由(a)知, } P(\bar{R}) = 1 - P(R) = 1 - \frac{2}{5} = \frac{3}{5}$$

$$(e) P(R \cup W) = \frac{\text{抽到一个红球或白球的可能方式数}}{\text{抽取一个球的所有可能方式数}} = \frac{6+4}{6+4+5} = \frac{10}{15} = \frac{2}{3}$$

$$\text{另解 由(c)知, } P(R \cup W) = P(\bar{B}) = 1 - P(B) = 1 - \frac{1}{3} = \frac{2}{3}.$$

注意, $P(R \cup W) = P(R) + P(W)$ (即 $\frac{2}{3} = \frac{2}{5} + \frac{4}{15}$). 这个例子可以说明对于互不相容事件 E_1 和 E_2 , 一般的结论 $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ 是正确的.

6.4 抛掷一个均匀的骰子两次. 求第一次出现 4, 5 或 6 且第二次出现 1, 2, 3 或 4 的概率.

解 设 E_1 为事件“第一次出现 4, 5 或 6”, E_2 为事件“第二次出现 1, 2, 3 或 4”. 将骰子第一次下落时的六种方式和第二次下落时的六种方式结合起来考虑, 一共有 $6 \times 6 = 36$ 种方式, 它们都是等可能的. E_1 发生的三种方式和 E_2 发生的四种方式结合起来, 一共有 $3 \times 4 = 12$ 种方式使得 E_1 和 E_2 同时发生或称 $E_1 E_2$ 发生. 则 $P(E_1 E_2) = 12/36 = 1/3$.

注意, $P(E_1 E_2) = P(E_1)P(E_2)$ (即 $\frac{1}{3} = \frac{3}{6} \cdot \frac{4}{6}$) 对于独立事件 E_1 和 E_2 成立.

6.5 从一副洗好的 52 张牌中抽取两张. 求在第一张牌(a)放回, (b)不放回, 两种取牌方式下两张都是 A 的概率.

解 设 E_1 为事件“第一次抽到 A”, E_2 为事件“第二次抽到 A”.

(a) 如果第一张牌放回去, 则 E_1 和 E_2 是独立事件. 那么

$$P(\text{两张都是 A}) = P(E_1 E_2) = P(E_1)P(E_2) = \frac{4}{52} \cdot \frac{4}{52} = \frac{1}{169}$$

(b) 第一张牌可在 52 张牌中抽取, 如不放回, 则第二张牌可在余下的 51 张中任取. 抽取两张牌有 52×51 种等可能的方法.

$$E_1 \text{ 可以有 4 种方式发生而 } E_1 \text{ 和 } E_2 \text{ 同时发生或者说 } E_1 E_2 \text{ 发生的方式有 } 4 \times 3 \text{ 种. 则 } P(E_1 E_2) = \frac{4 \times 3}{52 \times 51} = \frac{1}{221}.$$

注意, $P(E_2 | E_1) = P(\text{已知第一张是 A 时第二张也是 A}) = \frac{3}{51}$. 则上述结论说明当 E_1 和 E_2 不独立时, 一般的结论 $P(E_1 E_2) = P(E_1)P(E_2 | E_1)$ 成立.

6.6 从习题 6.3 中提到的盒子里连续取出三个球. 求在球(a)放回, (b)不放回. 两种取球方式下依次取到红球、白球和蓝球的概率.

解 设 R 为事件“第一次取到红球”, W 为事件“第二次取到白球”, B 为事件“第三次取到蓝球”, 要求 $P(RWB)$.

(a) 如果球是放回的, R , W 和 B 就是独立事件.

$$\begin{aligned} P(RWB) &= P(R)P(W)P(B) = \frac{6}{6+4+5} \cdot \frac{4}{6+4+5} \cdot \frac{5}{6+4+5} \\ &= \frac{6}{15} \cdot \frac{4}{15} \cdot \frac{5}{15} = \frac{8}{225} \end{aligned}$$

(b) 如果球是不放回的, R , W 和 B 不是独立事件.

$$\begin{aligned} P(RWB) &= P(R)P(W | R)P(B | WR) = \frac{6}{6+4+5} \cdot \frac{4}{5+4+5} \cdot \frac{5}{5+3+5} \\ &= \frac{6}{15} \cdot \frac{4}{14} \cdot \frac{5}{13} = \frac{4}{91} \end{aligned}$$

其中 $P(B | WR)$ 是已知第一次取出红球且第二次取出白球的条件下又取出蓝球的条件概率.

6.7 抛掷两次骰子, 求 4 至少出现一次的概率.

解 设 E_1 = 事件“第一次出现 4”, E_2 = 事件“第二次出现 4”. 则 $E_1 \cup E_2$ = 事件“第一次或第二次或两次同时出现 4” = 事件“至少一次出现 4”. 要求 $P(E_1 \cup E_2)$.

解法一 骰子下落两次有 $6 \times 6 = 36$ 种等可能的方式. 且

$$E_1 \text{ 发生而 } E_2 \text{ 不发生的方式数} = 5$$

$$E_2 \text{ 发生而 } E_1 \text{ 不发生的方式数} = 5$$

E_1 和 E_2 同时发生的方式数 = 1

则事件 E_1 或 E_2 至少一个发生的方式数目是 $5 + 5 + 1 = 11$, $P(E_1 \cup E_2) = \frac{11}{36}$.

解法二 由于 E_1 和 E_2 不是互不相容的, $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 E_2)$, 且 E_1 和 E_2 是独立的, $P(E_1 E_2) = P(E_1)P(E_2)$. 则 $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 E_2) = \frac{1}{6} + \frac{1}{6} - \frac{1}{6} \cdot \frac{1}{6} = \frac{11}{36}$.

解法三 $P(4 \text{ 至少出现一次}) + P(4 \text{ 一次也不出现}) = 1$
则

$$\begin{aligned} P(4 \text{ 至少出现一次}) &= 1 - P(4 \text{ 一次也不出现}) \\ &= 1 - P(\text{第一次不出现 } 4 \text{ 且第二次也不出现 } 4) \\ &= 1 - P(\bar{E}_1 \bar{E}_2) = 1 - P(\bar{E}_1)P(\bar{E}_2) \\ &= 1 - \frac{5}{6} \cdot \frac{5}{6} = \frac{11}{36} \end{aligned}$$

- 6.8** 一个袋子里装有 4 个白球和 2 个黑球; 另一个袋子里装有 3 个白球和 5 个黑球. 从每个袋子里取出一个球, 求下列事件发生的概率: (a) 两个球都是白球, (b) 两个球都是黑球, (c) 一个白球一个黑球.

解 设 W_1 为事件“从第一个袋子里取出的是白球”, W_2 为事件“从第二个袋子里取出的是白球”.

$$(a) P(W_1 W_2) = P(W_1)P(W_2) = \frac{4}{4+2} \cdot \frac{3}{3+5} = \frac{1}{4}$$

$$(b) P(\bar{W}_1 \bar{W}_2) = P(\bar{W}_1)P(\bar{W}_2) = \frac{2}{4+2} \cdot \frac{5}{3+5} = \frac{5}{24}$$

(c) 事件“一个白球一个黑球”相当于“第一个为白球第二个为黑球, 或第一个为黑球第二个为白球”, 即 $W_1 \bar{W}_2 \cup \bar{W}_1 W_2$. 由于 $W_1 \bar{W}_2$ 和 $\bar{W}_1 W_2$ 是互不相容的, 则有

$$\begin{aligned} P(W_1 \bar{W}_2 \cup \bar{W}_1 W_2) &= P(W_1 \bar{W}_2) + P(\bar{W}_1 W_2) \\ &= P(W_1)P(\bar{W}_2) + P(\bar{W}_1)P(W_2) \\ &= \frac{4}{4+2} \cdot \frac{5}{3+5} + \frac{2}{4+2} \cdot \frac{3}{3+5} = \frac{13}{24} \end{aligned}$$

另解 (c)要求的概率为 $1 - P(W_1 W_2) - P(\bar{W}_1 \bar{W}_2) = 1 - \frac{1}{4} - \frac{5}{24} = \frac{13}{24}$.

- 6.9** A 和 B 共下了 12 盘棋, 其中 A 赢了 6 盘, B 赢了 4 盘, 和了 2 盘. 他们决定再下三盘棋. 以前 12 盘棋的结果为经验概率, 求后 3 盘棋中下列事件发生的概率: (a) A 赢了 3 盘, (b) 和了 2 盘, (c) A 和 B 轮流获胜, (d) B 至少赢 1 盘.

解 设 A_1, A_2 和 A_3 分别表示 A 赢了第一盘, 第二盘和第三盘棋; B_1, B_2 和 B_3 分别表示 B 赢了第一盘、第二盘、第三盘棋; D_1, D_2 和 D_3 则分别表示和了第一盘, 第二盘和第三盘棋.

根据他们以往的战况(经验概率), 我们认为 $P(A \text{ 赢任一盘棋}) = \frac{6}{12} = \frac{1}{2}$, $P(B \text{ 赢任一盘棋}) = \frac{4}{12} = \frac{1}{3}$, $P(\text{和棋}) = \frac{2}{12} = \frac{1}{6}$.

$$(a) P(A \text{ 赢了 3 盘}) = P(A_1 A_2 A_3) = P(A_1)P(A_2)P(A_3) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

这里假定每盘棋的胜负结果是互相独立的, 这是合理的(除非选手们在心理上受到了其他胜负结果的影响).

$$\begin{aligned} (b) P(\text{和了 2 盘}) &= P(1, 2 \text{ 盘和或 } 1, 3 \text{ 盘和或 } 2, 3 \text{ 盘和}) \\ &= P(D_1 D_2 \bar{D}_3) + P(D_1 \bar{D}_2 D_3) + P(\bar{D}_1 D_2 D_3) \\ &= P(D_1)P(D_2)P(\bar{D}_3) + P(D_1)P(\bar{D}_2)P(D_3) + P(\bar{D}_1)P(D_2)P(D_3) \\ &= \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{5}{6} + \frac{1}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} + \frac{5}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{15}{216} = \frac{5}{72} \end{aligned}$$

$$\begin{aligned} (c) P(A \text{ 和 } B \text{ 轮流获胜}) &= P(\text{三盘棋结果依次为 } A \text{ 胜 } B \text{ 胜 } A \text{ 胜或 } B \text{ 胜 } A \text{ 胜 } B \text{ 胜}) \\ &= P(A_1 B_2 A_3 \cup B_1 A_2 B_3) = P(A_1 B_2 A_3) + P(B_1 A_2 B_3) \\ &= P(A_1)P(B_2)P(A_3) + P(B_1)P(A_2)P(B_3) \end{aligned}$$

$$= \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{3} = \frac{5}{36}$$

$$\begin{aligned} (d) P(B \text{ 至少赢一盘}) &= 1 - P(B \text{ 一盘也没赢}) \\ &= 1 - P(\bar{B}_1 \bar{B}_2 \bar{B}_3) = 1 - P(\bar{B}_1)P(\bar{B}_2)P(\bar{B}_3) \\ &= 1 - \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} = \frac{19}{27} \end{aligned}$$

概率分布

6.10 一个家庭有三个小孩,假定每个小孩是男是女是等可能的,求有不同个数男孩和女孩的概率.

解 设 B 为事件“小孩为男”, G 为事件“小孩为女”. 则根据等可能的假设, $P(B) = P(G) = \frac{1}{2}$. 在有三个孩子的家庭中, 下列互不相容事件以其相应的概率发生:

(a) 三个都是男孩(BBB):

$$P(BBB) = P(B)P(B)P(B) = \frac{1}{8}$$

这里我们假定前一个男孩的出生决不会影响到下一个男孩的出生, 即各事件是独立的.

(b) 三个都是女孩(GGG): 和(a)相对称

$$P(GGG) = \frac{1}{8}$$

(c) 两个男孩一个女孩($BBG \cup BGB \cup GBB$):

$$\begin{aligned} P(BBG \cup BGB \cup GBB) &= P(BBG) + P(BGB) + P(GBB) \\ &= P(B)P(B)P(G) + P(B)P(G)P(B) \\ &\quad + P(G)P(B)P(B) \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8} \end{aligned}$$

(d) 两个女孩一个男孩($GGB \cup GBG \cup BGG$): 和(c)相对称, 概率为 $\frac{3}{8}$.

如果我们以随机变量 X 表示三个孩子中男孩的人数, 则其概率分布如表 6.2 所示.

表 6.2

男孩人数 X	0	1	2	3
概率 $p(X)$	1/8	3/8	3/8	1/8

6.11 作图表示习题 6.10 中的分布.

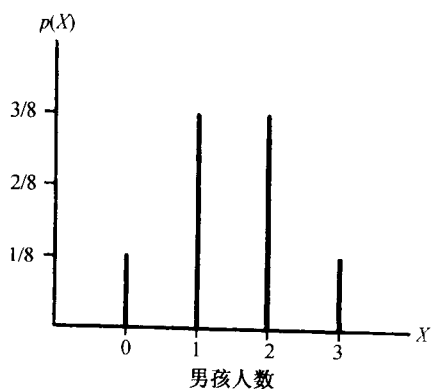


图 6-3

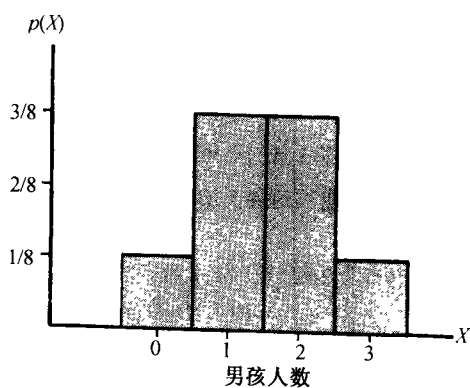


图 6-4

解 图 6-3 或 6-4 均可表示习题 6.10 中的分布. 注意, 图 6-4 中几个矩形的面积之和为 1, 这样的图称为**概率直方图**. 虽然 X 是离散的, 我们也把它当作连续型变量来考虑, 而且这种方法常常是非常有效的. 如果不想把变量当作连续的, 就要用类似于 6-3 这样的图.

6.12 一连续型随机变量只在 0 和 4 间取值, 密度函数为 $p(X) = \frac{1}{2} - aX$, a 是常数.

(a) 计算 a ;

(b) 求 $P(1 < X < 2)$.

解 (a) 如图 6-5, $p(X) = \frac{1}{2} - aX$ 的图像是一条直线. 为求得 a , 我们要认识到该直线的下方和 X 轴上方介于直线 $X=0$ 和 $X=4$ 之间的区域面积为 1. 当 $X=0$ 时 $p(X) = \frac{1}{2}$, 当 $X=4$ 时 $p(X) = \frac{1}{2} - 4a$. 然后选择 a 使得梯形的面积为 1. 梯形面积 $= \frac{1}{2} \times \text{高} \times \text{两底之和} = \frac{1}{2} \times 4 \times \left(\frac{1}{2} + \frac{1}{2} - 4a \right) = 2(1 - 4a) = 1$, 由此得 $1 - 4a = \frac{1}{2}$, $4a = \frac{1}{2}$, 则 $a = \frac{1}{8}$. 事实上, $\frac{1}{2} - 4a$ 就等于 0, 所以正确的图形应是图 6-6.

(b) 如图 6-6 所示, 要求的概率就是 $X=1$ 和 $X=2$ 之间的面积. 由 (a), $p(X) = \frac{1}{2} - \frac{1}{8}X$, 则 $p(1) = \frac{3}{8}$ 和 $p(2) = \frac{1}{4}$ 分别是对应于 $X=1$ 和 $X=2$ 的纵坐标. 要求的概率即是三角形的面积: $\frac{1}{2} \times 1 \times \left(\frac{3}{8} + \frac{1}{4} \right) = \frac{5}{16}$.

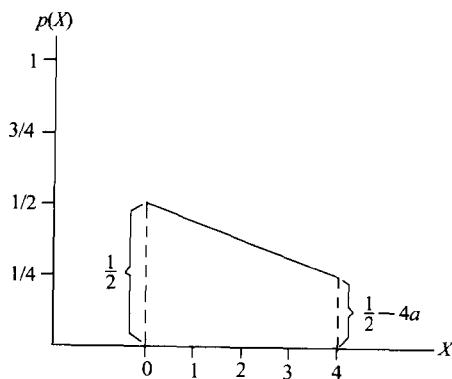


图 6-5

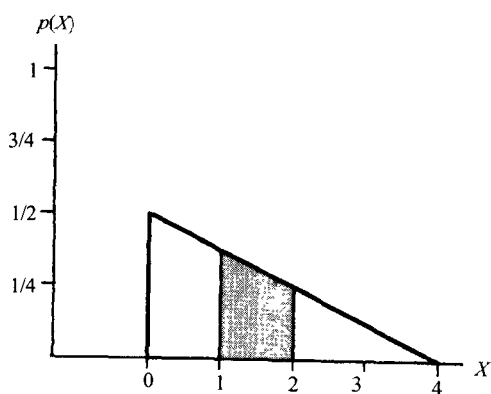


图 6-6

数学期望

6.13 一个人购买抽签售物的彩票, 中一等奖得 5000 美元和中二等奖得 2000 美元的概率分别是 0.001 和 0.003. 则一张彩票的合理价格是多少?

解 他的期望为 $5000 \text{ 美元} \times 0.001 + 2000 \text{ 美元} \times 0.003 = 5 \text{ 美元} + 6 \text{ 美元} = 11 \text{ 美元}$, 这是个合理的价格.

6.14 在一宗贸易投机中, 一位女士获利 300 美元的概率是 0.6, 亏本 100 美元的概率是 0.4. 求她的期望.

解 她的期望为 $300 \text{ 美元} \times 0.6 + (-100 \text{ 美元}) \times 0.4 = 180 \text{ 美元} - 40 \text{ 美元} = 140 \text{ 美元}$.

6.15 求表 6.3 表示的概率分布的 (a) $E(X)$, (b) $E(X^2)$, (c) $E[(X - E(X))^2]$.

表 6.3

X	8	12	16	20	24
$p(X)$	1/8	1/6	3/8	1/4	1/12

解 (a) $E(X) = \sum Xp(X) = 8 \times \frac{1}{8} + 12 \times \frac{1}{6} + 16 \times \frac{3}{8}$
 $+ 20 \times \frac{1}{4} + 24 \times \frac{1}{12} = 16$

这表示分布的均值.

(b) $E(X^2) = \sum X^2 p(X) = 8^2 \times \frac{1}{8} + 12^2 \times \frac{1}{6}$
 $+ 16^2 \times \frac{3}{8} + 20^2 \times \frac{1}{4} + 24^2 \times \frac{1}{12} = 276$

这表示二阶原点矩.

(c)

$$E[(X - E(X))^2] = \sum (X - E(X))^2 p(X) = (8 - 16)^2 \times \frac{1}{8}$$

$$+ (12 - 16)^2 \times \frac{1}{6} + (16 - 16)^2 \times \frac{3}{8}$$

$$+ (20 - 16)^2 \times \frac{1}{4} + (24 - 16)^2 \times \frac{1}{12} = 20$$

这表示分布的方差.

- 6.16 一个袋子中有 2 个白球和 3 个黑球. A, B, C 和 D 依次从袋子中取出一个球且不放回. 第一个取到白球的人将得到 10 美元奖金. 求 A, B, C 和 D 的期望.

解 因为袋子中只有 3 个黑球, 要想获胜必须第一次就取到白球. 分别用 A, B, C 和 D 表示事件“ A 获胜”, “ B 获胜”, “ C 获胜”和“ D 获胜”.

$$P(A \text{ 获胜}) = P(A) = \frac{2}{3+2} = \frac{2}{5}$$

则 A 的期望 $= \frac{2}{5} \times 10$ 美元 $= 4$ 美元.

$$P(A \text{ 失败 } B \text{ 获胜}) = P(\bar{A}B) = P(\bar{A})P(B|\bar{A}) = \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10}$$

则 B 的期望 $= 3$ 美元.

$$P(A, B \text{ 失败 } C \text{ 获胜}) = P(\bar{A}\bar{B}C) = P(\bar{A})P(\bar{B}|\bar{A})P(C|\bar{A}\bar{B})$$

$$= \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = \frac{1}{5}$$

则 C 的期望 $= 2$ 美元.

$$P(A, B \text{ 和 } C \text{ 失败 } D \text{ 获胜}) = P(\bar{A}\bar{B}\bar{C}D)$$

$$= P(\bar{A})P(\bar{B}|\bar{A})P(\bar{C}|\bar{A}\bar{B})P(D|\bar{A}\bar{B}\bar{C})$$

$$= \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} \cdot \frac{1}{1} = \frac{1}{10}$$

则 D 的期望 $= 1$ 美元.

检验: 4 美元 $+ 3$ 美元 $+ 2$ 美元 $+ 1$ 美元 $= 10$ 美元 且 $\frac{2}{5} + \frac{3}{10} + \frac{1}{5} + \frac{1}{10} = 1$.

排列

- 6.17 5 个不同颜色的球排成一行, 共有多少种排列方法?

解 我们把 5 个球安排在 5 个位置上: — — — — —. 第一个位置上可以安排 5 个弹球中的任何一个(即安排第一个位置有 5 种方法). 当第一个位置排定后, 安排第二个位置有 4 种方法. 安排第三个位置有 3 种方法, 第四个位置有 2 种方法. 安排最后一个位置时只剩下 1 种方法. 所以

$$5 \text{ 个球排成一行的排列数} = 5 \times 4 \times 3 \times 2 \times 1 = 5! = 120$$

一般地,

$$n \text{ 个不同物体排成一行的排列数} = n(n-1)(n-2)\cdots 1 = n!$$

这也称为从 n 个不同物体中一次取出 n 个的排列数, 记作 ${}_nP_n$.

- 6.18 一张长凳上有 4 个空位, 安排 10 个人坐. 共有多少种方法?

解 安排第一个位子可有 10 种方法, 第一个位子定下后安排第二个位子可有 9 种方法, 安排

第三个、第四个位子各有 8 种、7 种方法. 所以

$$10 \text{ 人坐 } 4 \text{ 个位子的排列数} = 10 \times 9 \times 8 \times 7 = 5040$$

一般地,

$$n \text{ 个不同物体中取出 } r \text{ 个的排列数} = n(n-1)\cdots(n-r+1)$$

这也称为从 n 个不同物体中一次取出 r 个的排列数, 记作 ${}_nP_r$, $P(n, r)$ 或 $P_{n,r}$. 注意到当 $r = n$ 时 ${}_nP_n = n!$, 如习题 6.17.

- 6.19 计算 (a) ${}_8P_3$, (b) ${}_6P_4$, (c) ${}_{15}P_1$, (d) ${}_3P_3$.

解 ✎ (a) ${}_8P_3 = 8 \times 7 \times 6 = 336$ (b) ${}_6P_4 = 6 \times 5 \times 4 \times 3 = 360$
(c) ${}_{15}P_1 = 15$ (d) ${}_3P_3 = 3 \times 2 \times 1 = 6$

- 6.20 5 位男士和 4 位女士坐成一排, 女士必须坐在偶数位上. 共有多少种可能的排法?

解 ✎ 男士共有 ${}_5P_5$ 种坐法, 女士共有 ${}_4P_4$ 种坐法; 将男士和女士的坐法结合起来, 则要求的排法数是 ${}_5P_5 \cdot {}_4P_4 = 5! \cdot 4! = 120 \times 24 = 2880$.

- 6.21 用 10 个数字 0, 1, 2, 3, \dots , 9 能组成多少个分别满足下列要求的四位数?

(a) 可以有重复数字, (b) 无重复数字, (c) 个位必须是 0 且无重复数字.

解 ✎ (a) 首位数字可以是 9 个数字中的任一个 (0 不可在首位). 第二、三、四位可以是 10 个数字中的任一个. 则可组成 $9 \times 10 \times 10 \times 10 = 9000$ 个数.

(b) 首位数字可以是 9 个数字中的任一个 (除 0 外). 第二位数字可以是其余 9 个数字中的任一个 (首位上的数字不可再用). 第三位数字可以是 8 个数字中的任一个 (前两位上的数字不可再用). 个位数字可以是 7 个数字中的任一个 (前三位上的数字不可再用). 则可组成 $9 \times 9 \times 8 \times 7 = 4536$ 个数.

另解 首位数字可以是 9 个数字中的任一个, 余下的三位数字共有 ${}_9P_3$ 种选法. 则可组成 $9 \cdot {}_9P_3 = 9 \times 9 \times 8 \times 7 = 4536$ 个数.

(c) 首位数字有 9 种选择, 第二位数字有 8 种选择, 第三位数字有 7 种选择. 则可组成 $9 \times 8 \times 7 = 504$ 个数.

另解 首位数字有 9 种选择, 中间两位数字有 ${}_8P_2$ 种选择. 则可组成 $9 \cdot {}_8P_2 = 9 \times 8 \times 7 = 504$ 个数.

- 6.22 书架上放有 4 本不同的数学书, 6 本不同的物理书和 2 本不同的化学书. 分别有多少种满足下列要求的排放方法? (a) 每门学科的书必须放在一起, (b) 只有数学书要放在一起.

解 ✎ (a) 数学书自身可有 ${}_4P_4 = 4!$ 种排法, 物理书有 ${}_6P_6 = 6!$ 种排法, 化学书有 ${}_2P_2 = 2!$ 种排法, 三种书有 ${}_3P_3 = 3!$ 种排法. 则要求的排列数 $= 4! \cdot 6! \cdot 2! \cdot 3! = 207360$.

(b) 将 4 本数学书看作一本大书, 那么我们共有 9 本书, 有 ${}_9P_9 = 9!$ 种排法. 在每一种排法中数学书都是放在一起的. 但数学书自身还有 ${}_4P_4 = 4!$ 种排法. 则要求的排列数 $= 9! \cdot 4! = 8709120$.

- 6.23 五个红色球, 两个白色球和三个蓝色球排成一行. 如果同种颜色的球不可区分, 共有多少种可能的排法?

解 ✎ 假设有 P 种不同的排法. P 乘以 (a) 5 个红球自身的排列数, 再乘以 (b) 2 个白球自身的排列数及 (c) 3 个蓝球自身的排列数 (即 P 乘以 $5! \cdot 2! \cdot 3!$), 就得到当 10 个球可以区分时的排列数 (即 $10!$). 则

$$5!2!3! \times P = 10!, \quad P = \frac{10!}{5!2!3!}$$

一般地, 如果 n 个物体中分别有 n_1, n_2, \dots, n_k 个是一样的, 不同的排列数为

$$\frac{n!}{n_1! n_2! \cdots n_k!}$$

其中 $n_1 + n_2 + \cdots + n_k = n$.

- 6.24 7 个人围着一张圆桌坐下, 分别有多少种满足下列要求的坐法? (a) 可任意坐, (b) 有 2 个人不能坐在一起.


解 ✎ (a) 让其中 1 个人随意坐下, 则其余 6 人有 $6! = 720$ 种坐法. 这就是 7 个人围成一圈的总的排列数.

(b) 将指定的 2 个人看作 1 个人, 则一共有 6 个人, 有 $5!$ 种坐法. 但是这 2 个人自身有 $2!$ 种坐法, 所以安排 7 个人围一圆桌坐下, 其中有 2 个指定的人坐在一起的方法有 $5! \cdot 2! = 240$ 种.

由 (a), 7 个人围一圆桌坐下但有 2 个人不能坐在一起的排列方法 $= 720 - 240 = 480$ 种.

组合

6.25 10 个物体分成两组, 每组分别包含 4 个和 6 个物体. 共有多少种分法?

解  这和分别有 4 个和 6 个是相同的 10 个物体的排列数是一样的. 由习题 6.23, 为


$$\frac{10!}{4!6!} = \frac{10 \times 9 \times 8 \times 7}{4!} = 210$$

这个问题相当于求从 10 个物体中选出 4 个的方法数 (或 10 个中选出 6 个), 而选择的顺序并不考虑.

一般地, 从 n 个物体中不考虑顺序地选出 r 个的选择方法数称为 n 中取 r 的**组合数**, 记为 $\binom{n}{r}$, 定义为

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{nPr}{r!}$$

6.26 计算 (a) $\binom{7}{4}$, (b) $\binom{6}{5}$, (c) $\binom{4}{4}$.

解  (a) $\binom{7}{4} = \frac{7!}{4!3!} = \frac{7 \times 6 \times 5 \times 4}{4!} = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} = 35$

(b) $\binom{6}{5} = \frac{6!}{5!1!} = \frac{6 \times 5 \times 4 \times 3 \times 2}{5!} = 6$ 或 $\binom{6}{5} = \binom{6}{1} = 6$

(c) $\binom{4}{4}$ 是 4 个物体全部选出的选择数, 而这种选择是惟一的, 则 $\binom{4}{4} = 1$. 注意, 若我们定义 $0! = 1$, 则从形式上有


$$\binom{4}{4} = \frac{4!}{4!0!} = 1$$

6.27 从 9 人中选出一个由 5 人组成的委员会, 共有多少种选法?

解 

$$\binom{9}{5} = \frac{9!}{5!4!} = \frac{9 \times 8 \times 7 \times 6 \times 5}{5!} = 126$$

6.28 从 5 位数学家和 7 位物理学家中选出一个由 2 位数学家和 3 位物理学家组成的委员会. 若 (a) 任何数学家和物理学家都可入选, (b) 一位指定的物理学家必须入选, (c) 两位指定的数学家不能入选, 各有多少种选法?

解  (a) 5 位数学家中选出 2 位共有 $\binom{5}{2}$ 种方法, 7 位物理学家中选出 3 位共有 $\binom{7}{3}$ 种方法, 一共可能有的选法数为

$$\binom{5}{2} \cdot \binom{7}{3} = 10 \times 35 = 350$$

(b) 5 位数学家中选出 2 位共有 $\binom{5}{2}$ 种方法, 6 位物理学家中选出另外的 2 位共有 $\binom{6}{2}$ 种方法, 一共可能有的选法数为

$$\binom{5}{2} \cdot \binom{6}{2} = 10 \times 15 = 150$$

(c) 3 位数学家中选出 2 位共有 $\binom{3}{2}$ 种方法, 7 位物理学家中选出 3 位共有 $\binom{7}{3}$ 种方法, 一共可能有的选法数为

$$\binom{3}{2} \cdot \binom{7}{3} = 3 \times 35 = 105$$

6.29 一个女孩有 5 支不同品种的花, 她可以扎成多少种不同的花束?

解 每支花会有 2 种处理方式: 选或不选. 将一支花的 2 种处理方式和其他花的 2 种处理方式结合起来, 5 支花的处理方式共有 2^5 种. 但 2^5 种包括一支花都没选中的情形. 因此要求的花束数目 $= 2^5 - 1 = 31$.

另解 她可以从 5 支花中选出 1 支, 2 支, \dots , 5 支扎成花束. 则要求的花束数目为

$$\binom{5}{1} + \binom{5}{2} + \binom{5}{3} + \binom{5}{4} + \binom{5}{5} = 5 + 10 + 10 + 5 + 1 = 31$$

一般地, 对于任意的正数 n

$$\binom{n}{1} + \binom{n}{2} + \binom{n}{3} + \dots + \binom{n}{n} = 2^n - 1$$

6.30 从 7 个辅音字母和 5 个元音字母中选出 4 个辅音字母和 3 个元音字母组成单词, 可组成多少单词? 单词不一定要有意义.

解 4 个不同的辅音字母有 $\binom{7}{4}$ 种选法, 3 个不同的元音字母可有 $\binom{5}{3}$ 种选法, 选出的 7 个不同的字母 (4 个辅音字母和 3 个元音字母) 自身有 $7P_7 = 7!$ 种排列方法. 则组成的单词数为

$$\binom{7}{4} \cdot \binom{5}{3} \cdot 7! = 35 \times 10 \times 5040 = 1\,764\,000$$

$n!$ 的 Stirling 逼近

6.31 计算 $50!$.

解 当 n 很大时, 我们有 $n! \approx \sqrt{2\pi n} n^n e^{-n}$, 则

$$50! \approx \sqrt{2\pi \cdot 50} \cdot 50^{50} e^{-50} = S$$

为求 S , 两边以 10 为底取对数, 则

$$\begin{aligned} \log S &= \log(\sqrt{100\pi} \cdot 50^{50} e^{-50}) = \frac{1}{2} \log 100 + \frac{1}{2} \log \pi + 50 \log 50 - 50 \log e \\ &= \frac{1}{2} \log 100 + \frac{1}{2} \log 3.142 + 50 \log 50 - 50 \log 2.718 \\ &= \frac{1}{2} \times 2 + \frac{1}{2} \times 0.4972 + 50 \times 1.6990 - 50 \times 0.4343 = 64.4846 \end{aligned}$$

由此得 $S = 3.05 \times 10^{64}$, S 有 65 位.

概率和组合分析

6.32 盒子里装有 8 个红球, 3 个白球和 9 个蓝球, 从中随机抽取 3 个球. 求下列事件发生的概率: (a) 3 个都是红球, (b) 3 个都是白球, (c) 2 个红球 1 个白球, (d) 至少有 1 个白球, (e) 每种颜色 1 个, (f) 顺次为红, 白, 蓝球.

解 (a) 解法一 设 R_1, R_2, R_3 分别表示事件“第一次、第二次、第三次取到的是红球”, 则 $R_1 R_2 R_3$ 表示取到的 3 个球都是红球.

$$P(R_1 R_2 R_3) = P(R_1)P(R_2 | R_1)P(R_3 | R_1 R_2) = \frac{8}{20} \cdot \frac{7}{19} \cdot \frac{6}{18} = \frac{14}{285}$$

解法二

$$\text{所求概率} = \frac{\text{8 个红球中取 3 个的方法数}}{\text{20 个球中取 3 个的方法数}} = \frac{\binom{8}{3}}{\binom{20}{3}} = \frac{14}{285}$$

(b) 同 (a) 的解法二

$$P(3 \text{ 个都是白球}) = \frac{\binom{3}{3}}{\binom{20}{3}} = \frac{1}{1140}$$

亦可用(a)中解法一的方法.

(c)

$$P(2 \text{ 个红球 } 1 \text{ 个白球}) = \frac{(8 \text{ 个红球中取 } 2 \text{ 个的方法数}) \times (3 \text{ 个白球中取 } 1 \text{ 个的方法数})}{20 \text{ 个球中取 } 3 \text{ 个的方法数}}$$

$$= \frac{\binom{8}{2} \binom{3}{1}}{\binom{20}{3}} = \frac{7}{95}$$

$$(d) P(1 \text{ 个白球也没有}) = \frac{\binom{17}{3}}{\binom{20}{3}} = \frac{34}{57}, \text{ 所以 } P(\text{至少有 } 1 \text{ 个白球}) = 1 - \frac{34}{57} = \frac{23}{57}$$

$$(e) P(\text{每种颜色 } 1 \text{ 个}) = \frac{\binom{8}{1} \binom{3}{1} \binom{9}{1}}{\binom{20}{3}} = \frac{18}{95}$$

$$(f) \text{ 由(e), } P(\text{顺次为红、白、蓝球}) = \frac{1}{3!} P(\text{每种颜色 } 1 \text{ 个}) = \frac{1}{6} \cdot \frac{18}{95} = \frac{3}{95}$$

另解 设 W_2 表示事件“第二次取到的是白球”, B_3 表示“第三次取到的是蓝球”, 则

$$\begin{aligned} P(R_1 W_2 B_3) &= P(R_1) P(W_2 | R_1) P(B_3 | R_1 W_2) \\ &= \frac{8}{20} \cdot \frac{3}{19} \cdot \frac{9}{18} = \frac{3}{95} \end{aligned}$$

6.33 从一副洗好的 52 张牌中抽出 5 张, 求抽到下列牌的概率:

(a) 4 张 A

(b) 4 张 A, 1 张 K

(c) 3 张 10, 2 张 J

(d) 9, 10, J, Q, K 各 1 张

(e) 其中有 3 张是同一花色的, 另 2 张是另一花色的

(f) 至少有 1 张 A

解 (a) $P(4 \text{ 张 A}) = \frac{\binom{4}{4} \binom{48}{1}}{\binom{52}{5}} = \frac{1}{54\,145}$

$$(b) P(4 \text{ 张 A, } 1 \text{ 张 K}) = \frac{\binom{4}{4} \binom{4}{1}}{\binom{52}{5}} = \frac{1}{649\,740}$$

$$(c) P(3 \text{ 张 } 10, 2 \text{ 张 } J) = \frac{\binom{4}{3} \binom{4}{2}}{\binom{52}{5}} = \frac{1}{108\,290}$$

$$(d) P(9, 10, J, Q, K \text{ 各 } 1 \text{ 张}) = \frac{\binom{4}{1} \binom{4}{1} \binom{4}{1} \binom{4}{1} \binom{4}{1}}{\binom{52}{5}} = \frac{64}{162\,435}$$

(e) 选定第一种花色时有 4 种选择, 选第二种花色时有 3 种选择, 则

$$P(3 \text{ 张同一花色, 另 } 2 \text{ 张为另一花色}) = \frac{4 \binom{13}{3} \cdot 3 \binom{13}{2}}{\binom{52}{5}} = \frac{429}{4165}$$

$$(f) P(1 \text{ 张 A 也没有}) = \frac{\binom{48}{5}}{\binom{52}{5}} = \frac{35\,673}{54\,145}, \text{ 则 } P(\text{至少有 } 1 \text{ 张 A}) = 1 - \frac{35\,673}{54\,145} = \frac{18\,482}{54\,145}$$

6.34 抛掷均匀的骰子 5 次, 求有 3 次得到 6 的概率.

解 用 5 个空格 — — — — — 表示 5 次抛掷的结果. 每一个空格里或者是 6 或者不是 6 ($\bar{6}$),

例如, 3 个是 6, 2 个不是 6, 可能为 $66\bar{6}\bar{6}$ 或 $6\bar{6}6\bar{6}$ 等.

$66\bar{6}\bar{6}$ 发生的概率为

$$\begin{aligned} P(66\bar{6}\bar{6}) &= P(6)P(6)P(\bar{6})P(\bar{6}) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \\ &= \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2 \end{aligned}$$

类似地, $P(6\bar{6}6\bar{6}) = \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2$, 所有的 3 个是 6, 2 个不是 6 的事件发生的概率都一样. 这样的事件共有 $\binom{5}{3} = 10$ 个, 且这些事件是互不相容的, 则所求概率为

$$P(66\bar{6}\bar{6} \text{ 或 } 6\bar{6}6\bar{6} \text{ 等等}) = \binom{5}{3} \cdot \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2 = \frac{125}{3888}$$

一般地, 如果 $p = P(E)$, $q = P(\bar{E})$, N 次独立试验中 E 发生 X 次的概率为 $\binom{N}{X} p^X q^{N-X}$

- 6.35** 某家工厂一台机器生产的螺栓中有 20% 是次品, 从这台机器某天的产品中随机选出 10 个螺栓, 求抽到下列产品的概率: (a) 2 个次品, (b) 2 个或 2 个以上次品, (c) 5 个以上次品.

解 (a) 同习题 6.34

$$P(2 \text{ 个次品}) = \binom{10}{2} 0.2^2 \cdot 0.8^8 = 45 \times 0.04 \times 0.1678 = 0.3020$$

(b)

$$\begin{aligned} P(2 \text{ 个或 } 2 \text{ 个以上次品}) &= 1 - P(\text{无次品}) - P(1 \text{ 个次品}) \\ &= 1 - \binom{10}{0} 0.2^0 \cdot 0.8^{10} - \binom{10}{1} 0.2^1 \cdot 0.8^9 \\ &= 1 - 0.8^{10} - 10 \times 0.2 \times 0.8^9 \\ &= 1 - 0.1074 - 0.2684 = 0.6242 \end{aligned}$$

(c)

$$\begin{aligned} P(5 \text{ 个以上次品}) &= P(6 \text{ 个次品}) + P(7 \text{ 个次品}) + P(8 \text{ 个次品}) \\ &\quad + P(9 \text{ 个次品}) + P(10 \text{ 个次品}) \\ &= \binom{10}{6} 0.2^6 \cdot 0.8^4 + \binom{10}{7} 0.2^7 \cdot 0.8^3 + \binom{10}{8} 0.2^8 \cdot 0.8^2 \\ &\quad + \binom{10}{9} 0.2^9 \cdot 0.8^1 + \binom{10}{10} 0.2^{10} \\ &= 0.00637 \end{aligned}$$

- 6.36** 从习题 6.35 提到的产品中选取 1000 组样品, 每组有 10 个螺栓, 我们预期从中能找出多少组满足下列条件的样品? (a) 2 个次品; (b) 2 个或 2 个以上次品; (c) 5 个以上次品.

解 (a) 由 6.35(a) 得, 期望值 $= 1000 \times 0.3020 = 302$.

(b) 由 6.35(b) 得, 期望值 $= 1000 \times 0.6242 \approx 624$.

(c) 由 6.35(c) 得, 期望值 $= 1000 \times 0.00637 \approx 6$.

样本空间和欧拉图

- 6.37** (a) 建立表示抛掷一对均匀骰子一次的结果的样本空间.
(b) 由样本空间求抛掷一对均匀骰子一次所得点数和为 7 或 11 的概率.

解 (a) 样本空间由图 6-7 中的点集构成. 每个点的第一个坐标是其中一个骰子上的数字, 第二个坐标是另一个骰子上的数字. 一共有 36 个点, 每个点对应的概率都为 $\frac{1}{36}$. 样本空间中所有点对应的概率和是 1.

(b) 图中 A 和 B 所指的点集分别表示事件“和为 7”与“和为 11”.

$$P(A) = A \text{ 中所有点的概率和 } = \frac{6}{36}$$

$$P(B) = B \text{ 中所有点的概率和 } = \frac{2}{36}$$

$$P(A \cup B) = A \text{ 和 } B \text{ 中所有点的概率和 } = \frac{6+2}{36} = \frac{8}{36} = \frac{2}{9}$$

注意, 本题中 $P(A \cup B) = P(A) + P(B)$, 这是因为 A 和 B 没有公共点 (即它们是互不相容事件).

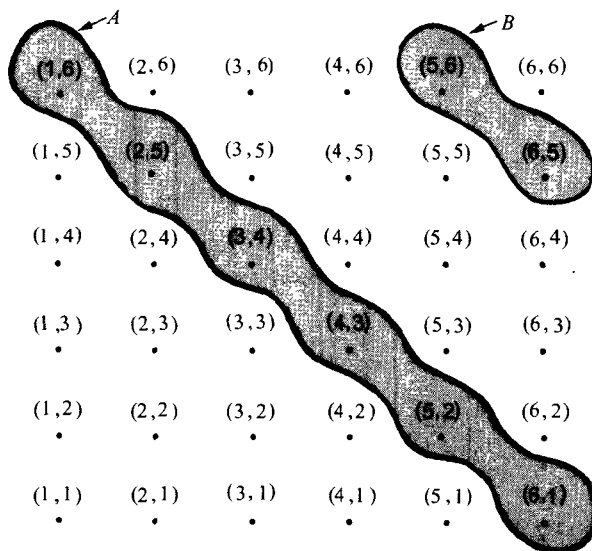


图 6-7

6.38 用样本空间理论证明:

(a) $P(A \cup B) = P(A) + P(B) - P(AB)$

(b) $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(BC) - P(AC) + P(ABC)$

证明 (a) 如图 6-8, 集合 A 和 B 的公共部分非空, 用 AB 表示. A 由 $A\bar{B}$ 和 AB 组成, 而 B 由 $B\bar{A}$ 和 AB 组成. $A \cup B$ 中所有点 = A 中所有点 + B 中所有点 - AB 中所有点. 因为一个事件或集合的概率就是集合中所有点所对应的概率之和, 所以有

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

另证 用 $A - AB$ 表示属于 A 但不属于 B 的点组成的集合 (即 $A\bar{B}$), 则 $A - AB$ 和 B 是互不相容的 (即它们无公共点), 而且 $P(A - AB) = P(A) - P(AB)$. 则

$$\begin{aligned} P(A \cup B) &= P(A - AB) + P(B) = P(A) - P(AB) + P(B) \\ &= P(A) + P(B) - P(AB) \end{aligned}$$

(b) 如图 6-9, A, B, C 为三个点集. 记号 $AB\bar{C}$ 表示属于 A 和 B 但不属于 C 的点组成的集合, 其他记号的含义同理可得.

将 A, B, C 中的点分成如图 6-9 所示的 7 个互不相容的集合. 所求的概率为

$$P(A \cup B \cup C) = P(AB\bar{C}) + P(BC\bar{A}) + P(C\bar{A}\bar{B})$$

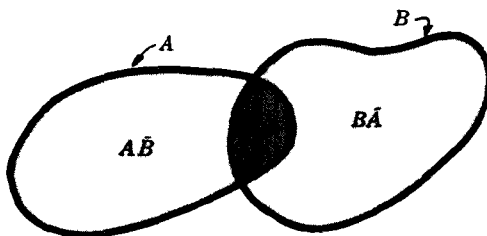


图 6-8

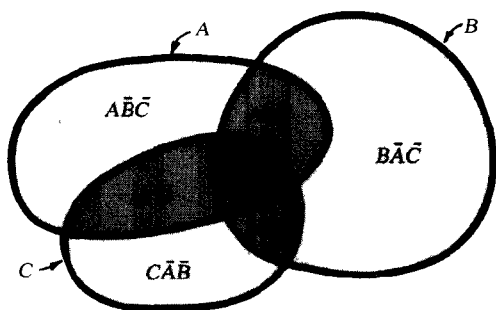


图 6-9

$$+ P(AB\bar{C}) + P(BC\bar{A}) + P(CA\bar{B}) + P(ABC)$$

为得到 $A\bar{B}\bar{C}$, 我们在 A 中除去 A 和 B 以及 A 和 C 的公共点, 但是这样一来 A, B 和 C 的公共点就去掉了两次. 因此 $A\bar{B}\bar{C} = A - AB - AC + ABC$ 且

$$P(A\bar{B}\bar{C}) = P(A) - P(AB) - P(AC) + P(ABC)$$

同理可得

$$P(\bar{B}\bar{C}A) = P(B) - P(BC) - P(BA) + P(BCA)$$

$$P(\bar{C}\bar{A}B) = P(C) - P(CA) - P(CB) + P(CAB)$$

$$P(BC\bar{A}) = P(BC) - P(ABC)$$

$$P(CA\bar{B}) = P(CA) - P(BCA)$$

$$P(AB\bar{C}) = P(AB) - P(CAB)$$

$$P(ABC) = P(ABC)$$

将 7 个方程相加并注意到 $P(AB) = P(BA)$ 等, 我们得到

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(BC) - P(AC) + P(ABC)$$

6.39 调查 500 名学生某学期选修代数、物理和统计学的情况, 数据如下:

代数	329	代数和物理	83
物理	186	代数和统计	217
统计	295	物理和统计	63

求满足下列条件的学生数:

- (a) 选了所有课
- (b) 选了代数未选统计
- (c) 选了物理未选代数
- (d) 选了统计未选物理
- (e) 选了代数或统计但未选物理
- (f) 选了代数但未选物理或统计

解 用 A 表示所有选修代数的学生的集合, (A) 表示这个集合的人数. 类似地, 用 (B) 表示选修物理的人数, 用 (C) 表示选修统计的人数. 则 $(A \cup B \cup C)$ 表示选修代数或物理或统计或几者兼有的人数, (AB) 表示既选代数又选物理的人数等. 同习题 6.38, 有

$$(A \cup B \cup C) = (A) + (B) + (C) - (AB) - (BC) - (AC) + (ABC)$$

(a) 将数据代入上式, 有

$$500 = 329 + 186 + 295 - 83 - 63 - 217 + (ABC)$$

得 $(ABC) = 53$, 这是三门课都选的人数. 注意, 一个学生三门课都选的(经验)概率为 $\frac{53}{500}$.

(b) 求解问题的一个简便方法是用欧拉图来表示属于各个集合的人数. 由于有 53 人选修了三门课, 我们可推出选修代数和统计但未选物理的人数是 $217 - 53 = 164$, 如图 6-10. 图中其他数目也可由已知数据求得.

由已有数据, 可得选修代数未选统计的人数 $= 329 - 217 = 112$; 或由图 6-10 得 $82 + 30 = 112$.

(c) 选物理未选代数的人数 $= 93 + 10 = 103$.

(d) 选统计未选物理的人数 $= 68 + 164 = 232$.

(e) 选代数或统计但未选物理的人数 $= 82 + 164 + 68 = 314$.

(f) 选代数但未选统计或物理的人数 $= 82$.

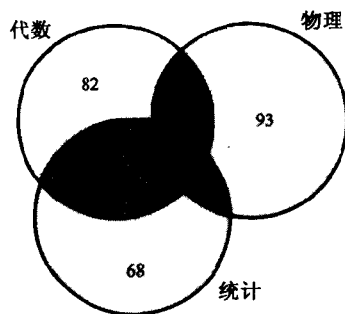


图 6-10

补充习题

概率的基本法则

- 6.40 求下列事件发生的概率 p 或其估计:
- 从一副洗好的牌中抽取一张, 抽到 K , A , 梅花 J 或方块 Q .
 - 抛掷一对均匀骰子一次, 点数和为 8.
 - 600 个螺栓中有 12 个是次品, 从中取出 1 个是正品.
 - 抛掷一对均匀骰子一次, 点数和为 7 或 11.
 - 抛掷一枚均匀硬币三次, 至少出现一次正面.
- 6.41 从一副洗好的牌中连续地抽三张牌. E_1 为事件“第一张是 K ”, E_2 为事件“第二张是 K ”, E_3 为事件“第三张是 K ”. 叙述下列记号的含义:
- $P(E_1 \bar{E}_2)$
 - $\bar{E}_1 \cup \bar{E}_2$
 - $\bar{E}_1 \bar{E}_2 \bar{E}_3$
 - $P(E_1 \cup E_2)$
 - $P(E_3 | E_1 \bar{E}_2)$
 - $P(E_1 E_2 \cup \bar{E}_2 E_3)$
- 6.42 从一个装有 10 个红球, 30 个白球, 20 个蓝球和 15 个黄球的盒子里随机抽取一个球. 求抽到下列球的概率: (a) 黄球或红球, (b) 不是红球或蓝球, (c) 不是蓝球, (d) 白球, (e) 红球, 白球或蓝球.
- 6.43 从习题 6.42 中的盒子里连续取两个球, 抽取是有放回的. 求抽到下列球的概率: (a) 两个白球, (b) 第一个是红球, 第二个是白球, (c) 两个都不是黄球, (d) 两个球或者是红球或者是白球, (e) 第二个不是蓝球, (f) 第一个是黄球, (g) 至少有一个是蓝球, (h) 至多有一个是红球, (i) 第一个是白球但第二个不是, (j) 只有一个是红球.
- 6.44 如果抽取是不放回的, 求解习题 6.43.
- 6.45 抛掷一对均匀骰子两次, 求 (a) 一次, (b) 至少一次, (c) 两次得到点数和为 7 的概率.
- 6.46 从一副洗好的 52 张牌中连续抽取 2 张, 求下列事件发生的概率: (a) 第一张不是梅花 10 也不是 A , (b) 第一张是 A 但第二张不是, (c) 至少有一张方块, (d) 两张不是同花的, (e) 至多有一张花牌 (J , Q , K), (f) 第二张不是花牌, (g) 第一张是花牌但第二张不是, (h) 两张是花牌或黑桃或两者都有.
- 6.47 盒子里装有标有 1 至 9 的 9 张票, 从中一次取出 3 张. 求它们依次是 (1) 奇数, 偶数, 奇数, (2) 偶数, 奇数, 偶数的概率.
- 6.48 A 和 B 下棋, A 对 B 的胜败比是 3:2. 若 A 和 B 下三盘棋, 下列事件发生的胜败比是多少? (a) A 至少赢两盘, (b) 前两盘 B 赢 A .
- 6.49 一个钱包里有 2 枚银币和 4 枚铜币, 另一个钱包里有 4 枚银币和 3 枚铜币. 从两个钱包中的一个里任取一枚硬币, 取出的是银币的概率是多少?
- 6.50 一位男士将再活 25 年的概率是 $\frac{3}{5}$, 他的妻子将再活 25 年的概率是 $\frac{2}{3}$. 求下列事件发生的概率: (a) 两人都能再活 25 年, (b) 只有男士可再活 25 年, (c) 只有他的妻子可再活 25 年, (d) 至少有 1 人可再活 25 年.
- 6.51 现有 800 个家庭, 每家 4 个孩子. 假定每个孩子是男是女是等可能的, 预计下列事件发生的百分比会是多少? (a) 有 2 个男孩 2 个女孩, (b) 至少有 1 个男孩, (c) 没有女孩, (d) 至多有 2 个女孩.

概率分布

- 6.52 如果随机变量 X 表示一个有 4 个孩子的家庭里男孩的个数 (见习题 6.51), (a) 列表表示 X 的概率分

布, (b) 画图表示 X 的概率分布.

- 6.53 连续型随机变量 X 只在 2 和 8 之内取值, 它的密度函数为 $a(X+3)$, 其中 a 是常数. (a) 计算 a , (b) 求 $P(3 < X < 5)$, (c) 求 $P(X \geq 4)$, (d) 求 $P(|X-5| < 0.5)$.
- 6.54 从一个装有 4 个红球和 6 个白球的罐子里无放回地取出 3 个球. 随机变量 X 表示取到的红球的个数. (a) 列表表示 X 的概率分布, (b) 画图表示 X 的概率分布.
- 6.55 求习题 6.54 中的 (a) $P(X=2)$, (b) $P(1 \leq X \leq 3)$, 并解释结果.

数学期望

- 6.56 某人参加游戏, 可以 0.2 的概率得到 25 美元, 可以 0.4 的概率得到 10 美元. 求参加游戏的合理价格是多少?
- 6.57 卖伞人下雨天每天可挣 30 美元, 晴天每天亏本 6 美元. 如果出现雨天的概率为 0.3, 那他收入的期望值是多少?
- 6.58 A 和 B 比赛抛掷硬币. 他们轮流抛掷一枚均匀硬币 3 次, 先使得硬币正面向上的人获胜. 如果 A 先抛且赌注是 20 美元, 那么每人应出多少钱比赛才是公平的?
- 6.59 求表 6.4 中的概率分布的 (a) $E(X)$, (b) $E(X^2)$, (c) $E[(X-E(X))^2]$, (d) $E(X^3)$.

表 6.4

X	-10	-20	30
$p(X)$	1/5	3/10	1/2

- 6.60 参考习题 6.54, 求 X 分布的 (a) 均值, (b) 方差, (c) 标准差, 并解释结果.
- 6.61 假设随机变量 X 以概率 p 取值 1, 以概率 $q=1-p$ 取值 0. 证明:
(a) $E(X)=p$, (b) $E[(X-E(X))^2]=pq$.
- 6.62 证明: (a) $E(2X+3)=2E(X)+3$, (b) $E[(X-E(X))^2]=E(X^2)-[E(X)]^2$.
- 6.63 X 和 Y 是两个具有相同分布的随机变量, 证明: $E(X+Y)=E(X)+E(Y)$.

排列

- 6.64 计算: (a) ${}_4P_2$, (b) ${}_7P_5$, (c) ${}_{10}P_3$.
- 6.65 求 n , 使得 ${}_{n+1}P_3 = {}_nP_4$ 成立.
- 6.66 沙发上只有 3 个位子, 有 5 个人想坐, 共有多少种安排方法?
- 6.67 7 本书放在书架上, 有多少种排列方法? 如果 (a) 任何排列都可以, (b) 某 3 本书必须放在一起, (c) 某 2 本书必须放在两端.
- 6.68 数字 1, 2, 3, ..., 9 可组成多少个没有重复数字的五位数? 如果 (a) 组成的五位数为奇数, (b) 每个数的前两位是偶数.
- 6.69 如果数字可以重复, 求解习题 6.68.
- 6.70 3 个 4, 4 个 2 和 2 个 3 可组成多少个不同的三位数?
- 6.71 3 位男士和 3 位女士围着圆桌坐下, 有多少种坐法? 如果 (a) 没有任何限制, (b) 某 2 位女士不能坐在一起, (c) 每位女士都要坐在两位男士中间.

组合

- 6.72 计算 (a) $\binom{7}{3}$, (b) $\binom{8}{4}$, (c) $\binom{10}{8}$.
- 6.73 求 n , 使得 $3\binom{n+1}{3} = 7\binom{n}{2}$ 成立.
- 6.74 10 个问题中选出 6 个, 有多少种选法?
- 6.75 从 8 位男士和 6 位女士中能选出多少个由 3 位男士和 4 位女士组成的不同的委员会?
- 6.76 从 6 位男士, 8 位女士, 4 个男孩和 5 个女孩中选出 2 位男士, 4 位女士, 3 个男孩和 3 个女孩, 有多少种方法? 如果 (a) 没有任何限制, (b) 某一位男士和某一位女士必须入选.
- 6.77 将 10 个人分组有多少种分法? 如果 (a) 分为两组, 一组 7 人, 一组 3 人, (b) 分为三组, 各有 4 人, 3 人

和 3 人.

- 6.78 从 5 位统计学家和 6 位经济学家中选出 3 位统计学家和 2 位经济学家组成一个委员会, 可有多少种方法? 如果 (a) 没有任何限制, (b) 某 2 位统计学家必须入选, (c) 某 1 位经济学家不能入选.
- 6.79 从单词 Tennessee 中选取 4 个字母, 求其 (a) 组合数, (b) 排列数.
- 6.80 证明 $1 - \binom{n}{1} + \binom{n}{2} - \binom{n}{3} + \cdots + (-1)^n \binom{n}{n} = 0$.

$n!$ 的 Stirling 逼近

- 6.81 100 个物体中选出 30 个排列, 有多少种方法?
- 6.82 证明: 当 n 很大时, $\binom{2n}{n} \approx 2^{2n} / \sqrt{\pi n}$.

杂题

- 6.83 从一副 52 张的牌中抽取 3 张. 求抽到下列牌的概率: (a) 两张 J 一张 K, (b) 三张同花的, (c) 三张不是同花的, (d) 至少两个 A.
- 6.84 抛掷一对均匀的骰子四次, 求至少两次点数和为 7 的概率.
- 6.85 如果一台机器生产的铆钉有 10% 是次品, 那么随机抽取的 5 个铆钉中 (a) 1 个次品也没有, (b) 有 1 个次品, (c) 至少有 2 个次品的概率是多少?
- 6.86 (a) 建立样本空间表示抛掷一枚均匀硬币两次的结果. 用 1 表示“正面”, 用 0 表示“反面”.
(b) 由样本空间求至少一次出现正面的概率.
(c) 能否建立一样本空间表示抛掷一枚硬币三次的结果? 若能, 用它求至多两次出现正面的概率.
- 6.87 某党三人 (A, B 和 C) 竞选三个职位, 下面是对 200 位选民的民意测验的结果:
28 人支持 A 和 B 122 人支持 B 或 C, 但不支持 A
98 人支持 A 或 B, 但不支持 C 64 人支持 C, 但不支持 A 或 B
42 人支持 B, 但不支持 A 或 C 14 人支持 A 和 C, 但不支持 B
则满足下列条件的有多少人? (a) 支持所有竞选者, (b) 支持 A, 不支持 B 或 C, (c) 支持 B, 不支持 A 或 C, (d) 支持 C, 不支持 A 或 B, (e) 支持 A 和 B, 不支持 C, (f) 只支持其中一人.
- 6.88 (a) 证明: 对任意事件 E_1 和 E_2 , $P(E_1 \cup E_2) \leq P(E_1) + P(E_2)$.
(b) 将 (a) 的结果推广.
- 6.89 E_1, E_2 和 E_3 是三个互不相容的事件且 $P(E_i) > 0, P(E_1 \cup E_2 \cup E_3) = 1$. A 为任一概率大于 0 的事件, 如果 $P(E_1), P(E_2), P(E_3)$ 和 $P(A|E_1), P(A|E_2), P(A|E_3)$ 已知, 证明:

$$P(E_1|A) = \frac{P(E_1)P(A|E_1)}{P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P(A|E_3)}$$

以及关于 $P(E_2|A)$ 和 $P(E_3|A)$ 的类似结果. 这就是贝叶斯法则或贝叶斯定理. 它在计算事件 A 发生条件下 E_1, E_2 或 E_3 发生的条件概率时是非常有用的. 这个结果可加以推广.

- 6.90 有三个不同的珠宝盒, 每个盒子有两个抽屉. 第一个盒子的每个抽屉里各有一块金表. 第二个盒子的每个抽屉里各有一块银表. 第三个盒子的一个抽屉里有一块金表, 另一个抽屉里有一块银表. 如果我们随机挑选一个盒子, 打开其中的一个抽屉, 发现里面有一块银表, 则另一抽屉里是金表的概率为多少? (提示: 利用习题 6.89 的结论)
- 6.91 求从数字 1, 2, 3, ..., 40 中选中 6 个无顺序的中奖数字的概率.
- 6.92 求解习题 6.91, 如果假设只要选中 (a) 5 个数字, (b) 4 个数字, (c) 3 个数字.
- 6.93 玩扑克时, 从一副 52 张的牌中给每人发五张. 求某人拿到下列牌的胜败比:
(a) 同花大顺 (同种花色的 A, K, Q, J 和 10);
(b) 同花顺 (同种花色的五张连续的牌, 如黑桃 3, 4, 5, 6 和 7);
(c) 四张一样的 (如四张 7);
(d) 三张一样, 另两张也一样 (如三张 K 和两张 10).
- 6.94 A 和 B 约定在下午 3 点和 4 点之间见面, 并说明每人等对方的时间不超过 10 分钟. 求他们能遇见的概率.
- 6.95 在长为 $a > 0$ 的线段上任取两点. 求形成的三条线段能构成三角形的概率.

第七章 二项分布, 正态分布和泊松分布

二项分布

在任一次试验中, 某事件发生的概率(称为**成功的概率**)为 p , 不发生的概率(称为**失败的概率**)为 $q = 1 - p$, 则在 N 次独立试验中该事件发生 X 次(即成功 X 次, 失败 $N - X$ 次)的概率为

$$p(X) = \binom{N}{X} p^X q^{N-X} = \frac{N!}{X!(N-X)!} p^X q^{N-X} \quad (1)$$

其中 $X = 0, 1, 2, \dots, N$, $N! = N(N-1)(N-2)\cdots 1$, $0! = 1$ (见习题 6.34 中定义).

例 1 抛掷一枚均匀硬币 6 次, 2 次出现正面的概率为(将 $N = 6$, $X = 2$, $p = q = \frac{1}{2}$ 代入 (1) 式)

$$\binom{6}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{6-2} = \frac{6!}{2!4!} \left(\frac{1}{2}\right)^6 = \frac{15}{64}$$

例 2 抛掷一枚均匀硬币 6 次, 至少 4 次出现正面的概率为

$$\begin{aligned} & \binom{6}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{6-4} + \binom{6}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^{6-5} + \binom{6}{6} \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^{6-6} \\ &= \frac{15}{64} + \frac{6}{64} + \frac{1}{64} = \frac{11}{32} \end{aligned}$$

离散型概率分布 (1) 常被称为**二项分布**, 因为当 $X = 0, 1, 2, \dots, N$ 时, 它逐项对应于**二项式公式**或**二项式展开式**:

$$(q + p)^N = q^N + \binom{N}{1} q^{N-1} p + \binom{N}{2} q^{N-2} p^2 + \cdots + p^N \quad (2)$$

中的各项, 其中 $1, \binom{N}{1}, \binom{N}{2}, \dots$ 称为**二项式系数**.

例 3

$$\begin{aligned} (q + p)^4 &= q^4 + \binom{4}{1} q^3 p + \binom{4}{2} q^2 p^2 + \binom{4}{3} q p^3 + p^4 \\ &= q^4 + 4q^3 p + 6q^2 p^2 + 4q p^3 + p^4 \end{aligned}$$

分布 (1) 是 James Bernoulli 在 17 世纪末发现的, 故也称为**伯努利分布**. 伯努利分布的一些性质列在表 7.1 中.

表 7.1 二项分布

均值	$\mu = Np$
方差	$\sigma^2 = Npq$
标准差	$\sigma = \sqrt{Npq}$
矩偏度系数	$\alpha_3 = \frac{q-p}{\sqrt{Npq}}$
矩峰度系数	$\alpha_4 = 3 + \frac{1-6pq}{Npq}$

例 4 抛掷一枚均匀硬币 100 次, 正面出现次数的均值是 $\mu = Np = 100 \times \frac{1}{2} = 50$, 这就是

抛掷一枚均匀硬币 100 次, 正面出现次数的期望值. 标准差是

$$\sigma = \sqrt{Npq} = \sqrt{100 \times \frac{1}{2} \times \frac{1}{2}} = 5.$$

正态分布

连续型概率分布的最重要的例子之一就是正态分布, 也称为高斯分布. 它由概率密度

$$Y = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} \quad (3)$$

定义. 其中 μ = 均值, σ = 标准差, $\pi = 3.14159\cdots$, $e = 2.71828\cdots$. 曲线(3)和 X 轴所围的总面积是 1, 因此, 曲线下方在 $X = a$ 和 $X = b$ 之间的面积表示 X 落在 a 和 b 之间的概率, 记为 $P(a < X < b)$, 其中 $a < b$.

将变量 X 标准化为 $Z = \frac{X-\mu}{\sigma}$, (3)式就变为所谓的标准形式:

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2} \quad (4)$$

此时我们称 Z 服从均值为 0 方差为 1 的标准正态分布. 图 7-1 是标准正态曲线的图像. 图像表明在 $Z = -1$ 和 $Z = +1$ 之间, $Z = -2$ 和 $Z = +2$ 之间, $Z = -3$ 和 $Z = +3$ 之间的面积分别为总面积的 68.27%, 95.45% 和 99.73%, 总面积是 1. 附录 II 中的表给出了曲线在 $Z = 0$ 和 $Z =$ 任意正数之间的面积. 利用曲线关于 $Z = 0$ 的对称性, 曲线在任意两平行于 y 轴的直线之间的面积都可由表得到.

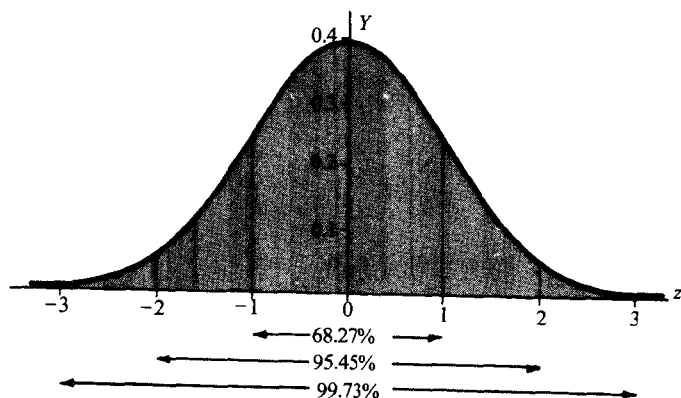


图 7-1

(3)式给出的正态分布的一些性质列在表 7.2 中.

表 7.2 正态分布

均值	μ
方差	σ^2
标准差	σ
矩偏度系数	$\alpha_3 = 0$
矩峰度系数	$\alpha_4 = 3$
平均偏差	$\sigma \sqrt{2/\pi} = 0.7979\sigma$

二项分布和正态分布的关系

如果 N 很大且 p 和 q 与 0 都不是很接近, 二项分布的标准化变量

$$Z = \frac{X - Np}{\sqrt{Npq}}$$

就近似于标准正态分布. 由表 7.1 和 7.2 可见, 当 N 越大时, 近似效果就越好, 在极限情形下结论是精确的. 当 N 越大时, 二项分布的偏度和峰度就越接近于正态分布的. 在实际情形中, 若 Np 和 Nq 都远远大于 5, 近似效果是非常好的.

泊松分布

离散型概率分布

$$p(X) = \frac{\lambda^X e^{-\lambda}}{X!} \quad X = 0, 1, 2, \dots \quad (5)$$

是 Siméon-Denis Poisson 在 19 世纪前叶发现的, 所以被称为**泊松分布**, 其中 $e = 2.71828 \dots$, λ 是一个给定的常数. $p(X)$ 的值可由附录 VIII 中的表(给出了不同的 λ 对应的 $e^{-\lambda}$ 值)或对数法求出.

泊松分布的一些性质列在表 7.3 中.

表 7.3 泊松分布

均值	$\mu = \lambda$
方差	$\sigma^2 = \lambda$
标准差	$\sigma = \sqrt{\lambda}$
矩偏度系数	$\alpha_3 = 1/\sqrt{\lambda}$
矩峰度系数	$\alpha_4 = 3 + 1/\lambda$

二项分布和泊松分布的关系

在二项分布(1)中, 如果 N 很大而事件发生的概率 p 接近于 0, 即 $q = 1 - p$ 接近于 1, 则称事件为**稀有事件**. 在实际情形中, 当试验的次数不少于 50 ($N \geq 50$), 而 Np 小于 5 时, 我们就认为它是稀有事件. 比较表 7.1 和表 7.3 可知此时, 二项分布非常近似于 $\lambda = Np$ 的泊松分布(5). 这是因为将 $\lambda = Np$, $q \approx 1$ 和 $p \approx 0$ 代入表 7.1, 就可得到表 7.3.

既然二项分布和正态分布之间存在某种关系, 那么泊松分布和正态分布之间也应该有某种关系. 事实上, 当 λ 趋向于正无穷大时, 泊松分布的标准化变量 $(X - \lambda)/\sqrt{\lambda}$ 就近似于标准正态分布.

多项分布

如果互不相容的事件 E_1, E_2, \dots, E_k 分别以概率 p_1, p_2, \dots, p_k 发生且 $p_1 + p_2 + \dots + p_k = 1$, 则在 N 次独立试验中事件 E_1, E_2, \dots, E_k 分别发生 X_1, X_2, \dots, X_k 次的概率为

$$\frac{N!}{X_1! X_2! \dots X_k!} p_1^{X_1} p_2^{X_2} \dots p_k^{X_k} \quad (6)$$

其中 $X_1 + X_2 + \dots + X_k = N$. 这一分布是二项分布的推广, 而表达式(6)是**多项展开式** $(p_1 + p_2 + \dots + p_k)^N$ 的一般项, 所以被称为**多项分布**.

例 5 抛掷一均匀骰子 12 次, 得到 1, 2, 3, 4, 5 和 6 各两次的概率为

$$\frac{12!}{2!2!2!2!2!2!} \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^2 = \frac{1925}{559872} = 0.00344$$

N 次试验中 E_1, E_2, \dots, E_k 发生次数的期望值分别是 Np_1, Np_2, \dots, Np_k .

用样本的频率分布拟合理论分布

如果可以通过概率推理或其他方法得到关于总体分布的一些信息,就有可能用总体的样本频率分布来拟合总体的**理论分布**(也称为**模型分布**或**期望分布**).常用的一般方法包括利用样本的均值和标准差来估计总体的均值和标准差(见习题 7.31, 7.33 和 7.34).

我们常用 χ^2 (希腊字母)**检验**(见第 12 章)来检验理论分布的**拟合优度**.判断正态分布是否是已知数据的一个好的拟合,简便方法是用**正态曲线图纸**,有时也叫**概率图纸**(见习题 7.32).

习题及解答

二项分布

7.1 计算下列值:

(a) $5!$ (b) $\frac{6!}{2!4!}$ (c) $\binom{8}{3}$ (d) $\binom{7}{5}$ (e) $\binom{4}{4}$ (f) $\binom{4}{0}$

解 (a) $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

(b) $\frac{6!}{2!4!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1) \times (4 \times 3 \times 2 \times 1)} = \frac{6 \times 5}{2 \times 1} = 15$

(c) $\binom{8}{3} = \frac{8!}{3!(8-3)!} = \frac{8!}{3!5!} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1) \times (5 \times 4 \times 3 \times 2 \times 1)} = \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = 56$

(d) $\binom{7}{5} = \frac{7!}{5!2!} = \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(5 \times 4 \times 3 \times 2 \times 1) \times (2 \times 1)} = \frac{7 \times 6}{2 \times 1} = 21$

(e) 由定义 $0! = 1$, 则 $\binom{4}{4} = \frac{4!}{4!0!} = 1$

(f) $\binom{4}{0} = \frac{4!}{0!4!} = 1$

7.2 假设在人群中有 15% 的人是左撇子, 求在 50 个人中下列事件发生的概率: (a) 至多有 10 个左撇子, (b) 至少有 5 个左撇子, (c) 有 3 到 6 个左撇子, (d) 正好有 5 个左撇子. 要求用 Minitab 求解.

解 (a) Minitab 的输出结果如下所示. 输入命令 cdf 10; 二项分布的子命令 $n = 50$ 和 $p = .15$, 即可得到要求的概率. 50 人中至多有 10 个左撇子的概率是 0.8801.

MTB > cdf 10;

SUBC > binomial $n = 50$ $p = .15$.

Cumulative Distribution Function

Binomial with $n = 50$ and $p = 0.150000$

x	P(X ≤ x)
10.0	0.8801

(b) Minitab 的输出结果如下所示. 至少有 5 个左撇子的逆事件是至多有 4 个左撇子. 由 $P(\text{事件}) = 1 - P(\text{逆事件})$, 有 $P(X \geq 5) = 1 - P(X \leq 4) = 1 - 0.1121 = 0.8879$.

MTB > cdf 4;

SUBC > binomial $n = 50$ $p = .15$.

Cumulative Distribution Function

Binomial with $n = 50$ and $p = 0.150000$

x	P(X ≤ x)
4.00	0.1121

(c) Minitab 的输出结果如下所示. $P(3 \leq X \leq 6) = P(X \leq 6) - P(X \leq 2) = 0.3613 - 0.0142 = 0.3471$.

MTB > cdf 6;

SUBC > binomial $n = 50$ $p = .15$.

Cumulative Distribution Function

Binomial with $n = 50$ and $p = 0.150000$

x	P(X ≤ x)
6.00	0.3613

MTB > cdf 2;

SUBC > binomial $n = 50$ $p = .15$.

Cumulative Distribution Function

Binomial with $n = 50$ and $p = 0.150000$

x	P(X ≤ x)
2.00	0.0142

(d) Minitab 的输出结果如下所示. 由结果可见 $P(X = 5) = 0.1072$.

MTB > pdf 5;

SUBC > binomial $n = 50$ $p = .15$.

Probability Density Function

Binomial with $n = 50$ and $p = 0.150000$

x	P(X = x)
5.00	0.1072

7.3 抛掷一均匀骰子 5 次, 求 3 出现 (a) 0 次, (b) 1 次, (c) 2 次, (d) 3 次, (e) 4 次的概率.

解 抛掷 1 次, 3 出现的概率 $p = \frac{1}{6}$, 不出现的概率 $q = 1 - p = \frac{5}{6}$, 则

$$(a) P(3 \text{ 出现 } 0 \text{ 次}) = \binom{5}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^5 = 1 \times 1 \times \left(\frac{5}{6}\right)^5 = \frac{3125}{7776}$$

$$(b) P(3 \text{ 出现 } 1 \text{ 次}) = \binom{5}{1} \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^4 = 5 \times \frac{1}{6} \times \left(\frac{5}{6}\right)^4 = \frac{3125}{7776}$$

$$(c) P(3 \text{ 出现 } 2 \text{ 次}) = \binom{5}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3 = 10 \times \frac{1}{36} \times \frac{125}{216} = \frac{625}{3888}$$

$$(d) P(3 \text{ 出现 } 3 \text{ 次}) = \binom{5}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2 = 10 \times \frac{1}{216} \times \frac{25}{36} = \frac{125}{3888}$$

$$(e) P(3 \text{ 出现 } 4 \text{ 次}) = \binom{5}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^1 = 5 \times \frac{1}{1296} \times \frac{5}{6} = \frac{25}{7776}$$

$$(f) P(3 \text{ 出现 } 5 \text{ 次}) = \binom{5}{5} \left(\frac{1}{6}\right)^5 \left(\frac{5}{6}\right)^0 = 1 \times \frac{1}{7776} \times 1 = \frac{1}{7776}$$

注意, 这些概率是下面的二项展开式的一般项:

$$\begin{aligned} \left(\frac{5}{6} + \frac{1}{6}\right)^5 &= \left(\frac{5}{6}\right)^5 + \binom{5}{1} \left(\frac{5}{6}\right)^4 \left(\frac{1}{6}\right)^1 + \binom{5}{2} \left(\frac{5}{6}\right)^3 \left(\frac{1}{6}\right)^2 \\ &\quad + \binom{5}{3} \left(\frac{5}{6}\right)^2 \left(\frac{1}{6}\right)^3 + \binom{5}{4} \left(\frac{5}{6}\right) \left(\frac{1}{6}\right)^4 + \left(\frac{1}{6}\right)^5 = 1 \end{aligned}$$

7.4 写出 (a) $(q + p)^4$, (b) $(q + p)^6$ 的二项展开式.

解 (a)

$$\begin{aligned} (q + p)^4 &= q^4 + \binom{4}{1} q^3 p + \binom{4}{2} q^2 p^2 + \binom{4}{3} q p^3 + p^4 \\ &= q^4 + 4q^3 p + 6q^2 p^2 + 4q p^3 + p^4 \end{aligned}$$

(b)

$$\begin{aligned} (q + p)^6 &= q^6 + \binom{6}{1} q^5 p + \binom{6}{2} q^4 p^2 + \binom{6}{3} q^3 p^3 + \binom{6}{4} q^2 p^4 + \binom{6}{5} q p^5 + p^6 \\ &= q^6 + 6q^5 p + 15q^4 p^2 + 20q^3 p^3 + 15q^2 p^4 + 6q p^5 + p^6 \end{aligned}$$

系数 1, 4, 6, 4, 1 和 1, 6, 15, 20, 15, 6, 1 分别被称为对应于 $N = 4$ 和 $N = 6$ 的二项式系数. 将 $N =$

0, 1, 2, 3, ... 时的系数写成如下所示的阵列, 我们就得到所谓的**帕斯卡三角**. 注意, 每一行的第一个数和最后一个数都是 1, 而其他的数则是它前一行的左边和右边的数之和.

$$\begin{array}{ccccccccccc}
 & & & & & & 1 & & & & & \\
 & & & & & 1 & & 1 & & & & \\
 & & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
 & 1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 & & 1
 \end{array}$$

- 7.5 在一个有 4 个孩子的家庭里, 假定男孩出生的概率是 $\frac{1}{2}$, 求下列事件发生的概率: (a) 至少有 1 个男孩, (b) 至少有 1 个男孩和 1 个女孩.

解 (a) $P(1 \text{ 个男孩}) = \binom{4}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = \frac{1}{4}$, $P(3 \text{ 个男孩}) = \binom{4}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1 = \frac{1}{4}$
 $P(2 \text{ 个男孩}) = \binom{4}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \frac{3}{8}$, $P(4 \text{ 个男孩}) = \binom{4}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^0 = \frac{1}{16}$
 则 $P(\text{至少有 1 个男孩}) = P(1 \text{ 个男孩}) + P(2 \text{ 个男孩}) + P(3 \text{ 个男孩}) + P(4 \text{ 个男孩})$
 $= \frac{1}{4} + \frac{3}{8} + \frac{1}{4} + \frac{1}{16} = \frac{15}{16}$

另解 $P(\text{至少有 1 个男孩}) = 1 - P(\text{没有男孩}) = 1 - \left(\frac{1}{2}\right)^4 = 1 - \frac{1}{16} = \frac{15}{16}$

(b) $P(\text{至少有 1 个男孩和 1 个女孩}) = 1 - P(\text{没有男孩}) - P(\text{没有女孩})$
 $= 1 - \frac{1}{16} - \frac{1}{16} = \frac{7}{8}$

- 7.6 在有 4 个孩子的 2000 个家庭中, 预计共有多少个家庭: (a) 至少有 1 个男孩, (b) 有 2 个男孩, (c) 有 1 个或 2 个女孩, (d) 没有女孩. 可参考习题 7.5(a).

解 (a) 预计有 1 个男孩的家庭数 $= 2000 \times \frac{15}{16} = 1875$

(b) 预计有 2 个男孩的家庭数 $= 2000 \cdot P(\text{有 2 个男孩}) = 2000 \times \frac{3}{8} = 750$

(c) $P(\text{有 1 个或 2 个女孩}) = P(\text{有 1 个女孩}) + P(\text{有 2 个女孩})$
 $= P(\text{有 1 个男孩}) + P(\text{有 2 个男孩})$
 $= \frac{1}{4} + \frac{3}{8} = \frac{5}{8}$

预计有 1 个或 2 个女孩的家庭数 $= 2000 \times \frac{5}{8} = 1250$

(d) 预计没有女孩的家庭数 $= 2000 \times \frac{1}{16} = 125$

- 7.7 如果一台机器生产的螺栓中有 20% 是次品, 从中随机抽取 4 个, 求下列事件发生的概率: (a) 有 1 个次品, (b) 没有次品, (c) 至多有 2 个次品.

解 一个螺栓是次品的概率为 $p = 0.2$, 不是次品的概率为 $q = 1 - p = 0.8$.

(a) $P(4 \text{ 个中有 1 个次品}) = \binom{4}{1} 0.2^1 \cdot 0.8^3 = 0.4096$

(b) $P(\text{没有次品}) = \binom{4}{0} 0.2^0 \cdot 0.8^4 = 0.4096$

(c) $P(\text{有 2 个次品}) = \binom{4}{2} 0.2^2 \cdot 0.8^2 = 0.1536$

则

$$\begin{aligned}
 P(\text{至多有 2 个次品}) &= P(\text{没有次品}) + P(1 \text{ 个次品}) + P(2 \text{ 个次品}) \\
 &= 0.4096 + 0.4096 + 0.1536 = 0.9728
 \end{aligned}$$

- 7.8 已知一名学生能毕业的概率是 0.4. 求在 5 名学生中下列事件发生的概率: (a) 无人能毕业, (b) 1 人能毕业, (c) 至少 1 人能毕业, (d) 都能毕业.

解 (a) $P(\text{无人能毕业}) = \binom{5}{0} 0.4^0 \cdot 0.6^5 = 0.07776$, 大约为 0.08

$$(b) P(1 \text{ 人能毕业}) = \binom{5}{1} 0.4^1 \cdot 0.6^4 = 0.2592, \text{ 大约为 } 0.26$$

$$(c) P(\text{至少 } 1 \text{ 人能毕业}) = 1 - P(\text{无人能毕业}) = 0.92224, \text{ 大约为 } 0.92$$

$$(d) P(\text{都能毕业}) = \binom{5}{5} 0.4^5 \cdot 0.6^0 = 0.01024, \text{ 大约为 } 0.01$$

7.9 抛掷一对骰子 6 次, 求得到点数和为 9 (a) 两次, (b) 至少两次的概率是多少?

解 第一个骰子的 6 种下落方式可和第二个骰子的 6 种下落方式结合起来考虑, 则一对骰子的下落方式一共有 $6 \times 6 = 36$ 种. 它们分别是: 第一个骰子出现 1, 第二个骰子也出现 1; 第一个骰子出现 1, 第二个骰子出现 2 等. 可分别记为 (1, 1), (1, 2) 等.

在 36 种方式中 (如果骰子是均匀的, 则它们的出现是等可能的), 点数和为 9 的情形有 4 种: (3, 6), (4, 5), (5, 4) 和 (6, 3). 则抛掷一对骰子一次, 点数和为 9 的概率是 $p = \frac{4}{36} = \frac{1}{9}$, 而和不为 9 的概率是 $q = 1 - p = \frac{8}{9}$.

$$(a) P(2 \text{ 次出现 } 9) = \binom{6}{2} \left(\frac{1}{9}\right)^2 \left(\frac{8}{9}\right)^{6-2} = \frac{61\,440}{531\,441}$$

$$\begin{aligned} (b) P(\text{至少 } 2 \text{ 次出现 } 9) &= P(2 \text{ 次出现 } 9) + P(3 \text{ 次出现 } 9) + P(4 \text{ 次出现 } 9) \\ &\quad + P(5 \text{ 次出现 } 9) + P(6 \text{ 次出现 } 9) \\ &= \binom{6}{2} \left(\frac{1}{9}\right)^2 \left(\frac{8}{9}\right)^4 + \binom{6}{3} \left(\frac{1}{9}\right)^3 \left(\frac{8}{9}\right)^3 + \binom{6}{4} \left(\frac{1}{9}\right)^4 \left(\frac{8}{9}\right)^2 \\ &\quad + \binom{6}{5} \left(\frac{1}{9}\right)^5 \left(\frac{8}{9}\right)^1 + \binom{6}{6} \left(\frac{1}{9}\right)^6 \left(\frac{8}{9}\right)^0 \\ &= \frac{61\,440}{531\,441} + \frac{10\,240}{531\,441} + \frac{960}{531\,441} \\ &\quad + \frac{48}{531\,441} + \frac{1}{531\,441} = \frac{72\,689}{531\,441} \end{aligned}$$

另解 $P(\text{至少出现 } 2 \text{ 次}) = 1 - P(\text{出现 } 0 \text{ 次}) - P(\text{出现 } 1 \text{ 次})$

$$\begin{aligned} &= 1 - \binom{6}{0} \left(\frac{1}{9}\right)^0 \left(\frac{8}{9}\right)^6 - \binom{6}{1} \left(\frac{1}{9}\right)^1 \left(\frac{8}{9}\right)^5 \\ &= \frac{72\,689}{531\,441} \end{aligned}$$

7.10 计算: (a) $\sum_{X=0}^N X p(X)$ 和 (b) $\sum_{X=0}^N X^2 p(X)$, 其中 $p(X) = \binom{N}{X} p^X q^{N-X}$.

解 (a) 因为 $q + p = 1$, 所以有

$$\begin{aligned} \sum_{X=0}^N X p(X) &= \sum_{X=1}^N X \frac{N!}{X!(N-X)!} p^X q^{N-X} \\ &= Np \sum_{X=1}^N \frac{(N-1)!}{(X-1)!(N-X)!} p^{X-1} q^{N-X} \\ &= Np(q + p)^{N-1} = Np \end{aligned}$$

$$\begin{aligned} (b) \sum_{X=0}^N X^2 p(X) &= \sum_{X=1}^N X^2 \frac{N!}{X!(N-X)!} p^X q^{N-X} \\ &= \sum_{X=1}^N [X(X-1) + X] \frac{N!}{X!(N-X)!} p^X q^{N-X} \\ &= \sum_{X=2}^N X(X-1) \frac{N!}{X!(N-X)!} p^X q^{N-X} \\ &\quad + \sum_{X=1}^N X \frac{N!}{X!(N-X)!} p^X q^{N-X} \\ &= N(N-1)p^2 \sum_{X=2}^N \frac{(N-2)!}{(X-2)!(N-X)!} p^{X-2} q^{N-X} + Np \\ &= N(N-1)p^2(q + p)^{N-2} + Np \\ &= N(N-1)p^2 + Np \end{aligned}$$

注意, (a) 和 (b) 的结果分别是 X 和 X^2 的期望, 记为 $E(X)$ 和 $E(X^2)$ (见第六章).

7.11 如果一个变量服从二项分布, 求它的 (a) 均值 μ , (b) 方差 σ^2 .

解 (a) 由习题 7.10(a)

$$\mu = \text{变量的均值} = \sum_{X=0}^N Xp(X) = Np$$

(b) 由 $\mu = Np$ 及习题 7.10 的结果

$$\begin{aligned}\sigma^2 &= \sum_{X=0}^N (X - \mu)^2 p(X) = \sum_{X=0}^N (X^2 - 2\mu X + \mu^2) p(X) \\ &= \sum_{X=0}^N X^2 p(X) - 2\mu \sum_{X=0}^N X p(X) + \mu^2 \sum_{X=0}^N p(X) \\ &= N(N-1)p^2 + Np - 2Np \cdot Np + (Np)^2 \cdot 1 \\ &= Np - Np^2 = Np(1-p) = Npq\end{aligned}$$

由此得二项分布的标准差是 $\sigma = \sqrt{Npq}$.

另解 由习题 6.62(b)

$$\begin{aligned}E[(X - E(X))^2] &= E(X^2) - [E(X)]^2 = N(N-1)p^2 + Np - N^2p^2 \\ &= Np - Np^2 = Npq\end{aligned}$$

7.12 如果一个螺栓是次品的概率是 0.1, 求 400 个螺栓中次品数所服从的分布的 (a) 均值, (b) 标准差.

解 (a) 均值 $Np = 400 \times 0.1 = 40$, 即我们可预计有 40 个是次品.

(b) 方差 $Npq = 400 \times 0.1 \times 0.9 = 36$, 则标准差是 $\sqrt{36} = 6$.

7.13 求习题 7.12 中的 (a) 矩偏度系数和 (b) 矩峰度系数.

解 (a) 矩偏度系数 $= \frac{q-p}{\sqrt{Npq}} = \frac{0.9-0.1}{6} = 0.133$

因为系数为正, 所以分布向右偏.

(b) 峰度矩系数 $= 3 + \frac{1-6pq}{Npq} = 3 + \frac{1-6 \times 0.1 \times 0.9}{36} = 3.01$

这个分布相对于正态分布来说略有点尖峰 (见第五章).

正态分布

7.14 一次数学测验的均分是 72, 标准差是 15. 求 (a) 60 分, (b) 93 分, (c) 72 分的标准分数.

解 (a) $z = \frac{X - \bar{X}}{s} = \frac{60 - 72}{15} = -0.8$

(b) $z = \frac{X - \bar{X}}{s} = \frac{93 - 72}{15} = 1.4$

(c) $z = \frac{X - \bar{X}}{s} = \frac{72 - 72}{15} = 0$

7.15 参见习题 7.14, 求对应于标准分数 (a) -1, (b) 1.6 的分数.

解 (a) $X = \bar{X} + zs = 72 + (-1) \times 15 = 57$

(b) $X = \bar{X} + zs = 72 + 1.6 \times 15 = 96$

7.16 假定一名棒球运动员在他的职业生涯中参赛的场数服从均值为 1500, 标准差为 350 的正态分布. 用 Minitab 解答下列问题: (a) 参赛少于 750 场的百分比是多少? (b) 参赛多于 2000 场的百分比是多少? (c) 求参赛场数的第 90 个百分位数.

解 (a) 由下面的 Minitab 的结果, 可得 $P(X < 750) = 0.0161$, 或参赛少于 750 场的占 1.61%.

MTB > cdf 750;

SUBC > normal mean = 1500 sd = 350.

Cumulative Distribution Function

Normal with mean = 1500.00 and standard deviation = 350.000

x	P(X ≤ x)
750.0000	0.0161

(b) 由下面的 Minitab 的结果, 可得 $P(X < 2000) = 0.9234$. 由此得 $P(X > 2000) = 1 - P(X < 2000) = 1 - 0.9234 = 0.0766$. 因此参赛多于 2000 场的占 7.66%.

MTB > cdf 2000;

SUBC > normal mean = 1500 sd = 350.

Cumulative Distribution Function

Normal with mean = 1500.00 and standard deviation = 350.000

x	P(X ≤ x)
2.00E+03	0.9234

(c) Minitab 给出的第 90 个百分位数是 1.95E+03 或 1950 场.

MTB > invcdf .90;

SUBC > normal mean = 1500 sd = 350.

Inverse Cumulative Distribution Function

Normal with mean = 1500.00 and standard deviation = 350.000

P(X ≤ x)	x
0.9000	1.95E+03

7.17 利用附录 II, 求分别对应于图 7-2 (a) 至 (g) 的下列情形下的正态曲线下方的面积:

- (a) 在 $z=0$ 和 $z=1.2$ 之间
- (b) 在 $z=-0.68$ 和 $z=0$ 之间
- (c) 在 $z=-0.46$ 和 $z=2.21$ 之间
- (d) 在 $z=0.81$ 和 $z=1.94$ 之间
- (e) 在 $z=-0.6$ 左方
- (f) 在 $z=-1.28$ 右方
- (g) 在 $z=2.05$ 右方, $z=-1.44$ 左方

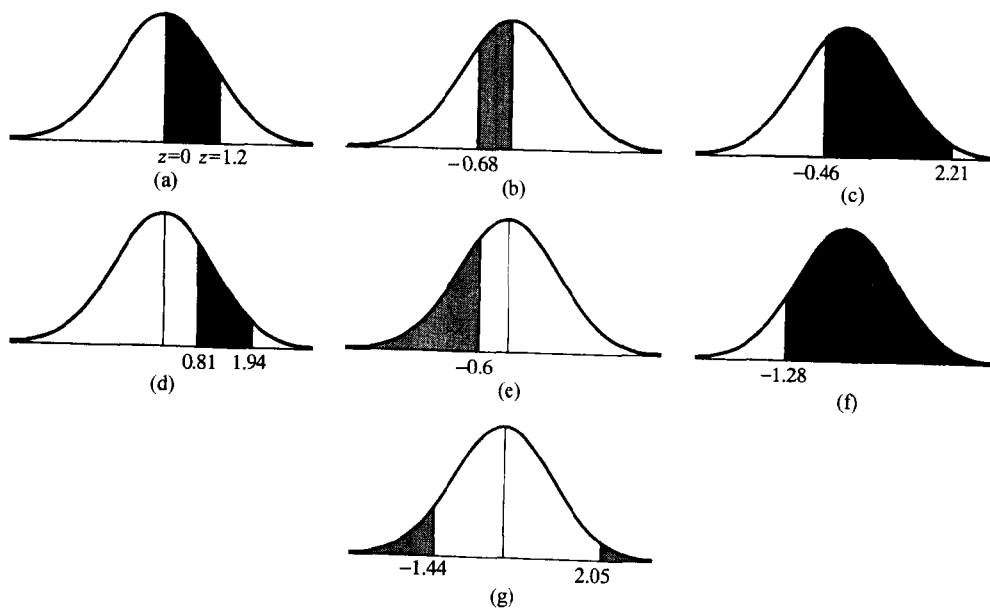


图 7-2

解 (a) 在附录 II 中, 沿着标有 z 的一栏向下直到找到 1.2; 再向右直到找到标有 0 的一列. 结果是 0.3849, 即为所要求的面积, 表示 z 在 0 和 1.2 之间的概率, 记为 $P(0 \leq z \leq 1.2)$.

(b) 所要求的面积即为在 $z=0$ 和 $z=0.68$ 之间的面积 (由对称性). 为求此面积, 在附录 II 中, 沿着标有 z 的一栏向下直到找到 0.6; 再向右直到找到标有 8 的一列. 结果是 0.2517, 即为所要求的面积, 表示 z 在 -0.68 和 0 之间的概率, 记为 $P(-0.68 \leq z \leq 0)$.

(c) 要求的面积 = ($z = -0.46$ 和 $z = 0$ 之间的面积) + ($z = 0$ 和 $z = 2.21$ 之间的面积)
= ($z = 0$ 和 $z = 0.46$ 之间的面积) + ($z = 0$ 和 $z = 2.21$ 之间的面积)

$$= 0.1772 + 0.4864 = 0.6636$$

$$\begin{aligned} \text{(d) 要求的面积} &= (z=0 \text{ 和 } z=1.94 \text{ 之间的面积}) - (z=0 \text{ 和 } z=0.81 \text{ 之间的面积}) \\ &= 0.4738 - 0.2910 = 0.1828 \end{aligned}$$

$$\begin{aligned} \text{(e) 要求的面积} &= (z=0 \text{ 左方的面积}) - (z=-0.6 \text{ 和 } z=0 \text{ 之间的面积}) \\ &= (z=0 \text{ 左方的面积}) - (z=0 \text{ 和 } z=0.6 \text{ 之间的面积}) \\ &= 0.5 - 0.2258 = 0.2742 \end{aligned}$$

$$\begin{aligned} \text{(f) 要求的面积} &= (z=-1.28 \text{ 和 } z=0 \text{ 之间的面积}) + (z=0 \text{ 右方的面积}) \\ &= 0.3997 + 0.5 = 0.8997 \end{aligned}$$

$$\begin{aligned} \text{(g) 要求的面积} &= \text{总面积} - (z=-1.44 \text{ 和 } z=0 \text{ 之间的面积}) \\ &\quad - (z=0 \text{ 和 } z=2.05 \text{ 之间的面积}) \\ &= 1 - 0.4251 - 0.4798 = 1 - 0.9049 = 0.0951 \end{aligned}$$

7.18 分别求图 7-3 中如下给定的面积所对应的 z 值.“面积”指的是正态曲线下方的面积.

- (a) 0 和 z 之间的面积是 0.3770;
 (b) z 左方的面积是 0.8621;
 (c) -1.5 和 z 之间的面积是 0.0217.

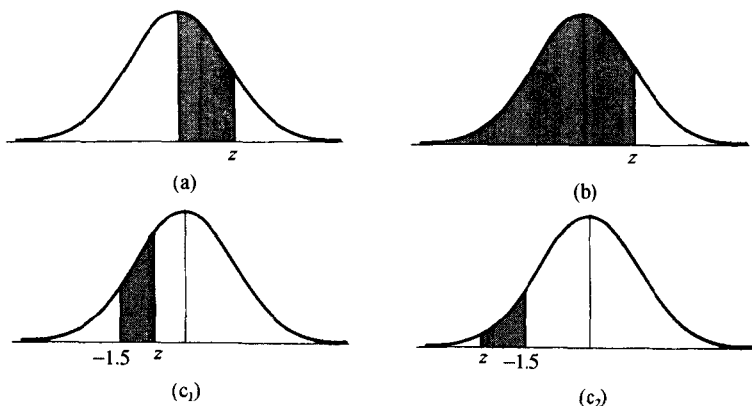


图 7-3

解 (a) 在附录 II 中, 值 0.3770 位于标有 1.1 的行的右方和标有 6 的列的下方, 则要求的 $z = 1.16$. 由对称性, $z = -1.16$ 是另一解, 则 $z = \pm 1.16$.

(b) 既然面积大于 0.5, z 必为正数. 0 和 z 之间的面积 $= 0.8621 - 0.5 = 0.3621$, 则 $z = 1.09$.

(c) 如果 z 为正数, 面积应大于 -1.5 和 0 之间的面积 0.4332, 因此 z 必为负数.

情形 1 [z 为负数, 但在 -1.5 的右方, 见图 7-3(c₁)]

-1.5 和 z 之间的面积 $= (-1.5 \text{ 和 } 0 \text{ 之间的面积}) - (0 \text{ 和 } z \text{ 之间的面积})$, 即 $0.0217 = 0.4332 - (0 \text{ 和 } z \text{ 之间的面积})$. 则 0 和 z 之间的面积 $= 0.4332 - 0.0217 = 0.4115$, 因此 $z = -1.35$.

情形 2 [z 为负数, 但在 -1.5 的左方, 见图 7-3(c₂)]

z 和 -1.5 之间的面积 $= (z \text{ 和 } 0 \text{ 之间的面积}) - (-1.5 \text{ 和 } 0 \text{ 之间的面积})$, 即 $0.0217 = (0 \text{ 和 } z \text{ 之间的面积}) - 0.4332$. 则 0 和 z 之间的面积 $= 0.0217 + 0.4332 = 0.4549$, 由线性插值 $z = -1.694$, 或精确到 $z = -1.69$.

7.19 求正态曲线在 (a) $z = 0.84$, (b) $z = -1.27$, (c) $z = -0.05$ 处的纵坐标.

解 (a) 在附录 I 中, 沿着标有 z 的一列向下找到值 0.8, 然后向右找到标有 4 的列. 值 0.2803 就是要求的纵坐标.

(b) 由对称性: (在 $z = -1.27$ 处的纵坐标) $=$ (在 $z = 1.27$ 处的纵坐标) $= 0.1781$.

(c) (在 $z = -0.05$ 处的纵坐标) $=$ (在 $z = 0.05$ 处的纵坐标) $= 0.3984$.

7.20 某大学 500 名男同学的平均重量是 151 磅, 标准差是 15 磅. 假设重量是服从正态分布的, 求有多少同学的重量满足下列要求: (a) 在 120 磅和 155 磅之间, (b) 大于 185 磅.

解 (a) 假设记录每个人的重量时都是四舍五入的, 则记录为 120 磅和 155 磅之间的重量事实

上可能是 119.5 磅到 155.5 磅之间的任何值:

$$119.5 \text{ 的标准值} = \frac{119.5 - 151}{15} = -2.10$$

$$155.5 \text{ 的标准值} = \frac{155.5 - 151}{15} = 0.30$$

如图 7-4(a)

$$\begin{aligned} \text{要求的学生比例} &= (z = -2.10 \text{ 和 } z = 0.30 \text{ 之间的面积}) \\ &= (z = -2.10 \text{ 和 } z = 0 \text{ 之间的面积}) \\ &\quad + (z = 0 \text{ 和 } z = 0.30 \text{ 之间的面积}) \\ &= 0.4821 + 0.1179 = 0.6000 \end{aligned}$$

则重量在 120 磅和 155 磅之间的学生数 $= 500 \times 0.6000 = 300$.

(b) 记录重量大于 185 磅的学生至少重 185.5 磅.

$$185.5 \text{ 的标准值} = \frac{185.5 - 151}{15} = 2.30$$

如图 7-4(b)

$$\begin{aligned} \text{要求的学生比例} &= (z = 2.30 \text{ 右方的面积}) \\ &= (z = 0 \text{ 右方的面积}) - (z = 0 \text{ 和 } z = 2.30 \text{ 之间的面积}) \\ &= 0.5 - 0.4893 = 0.0107 \end{aligned}$$

则重量大于 185 磅的学生数 $= 500 \times 0.0107 = 5$.

如果用 W 表示随机抽到的一名学生的重量, 我们可用概率把上面的结果归纳如下:

$$P(119.5 \leq W \leq 155.5) = 0.6000, \quad P(W \geq 185.5) = 0.0107$$



图 7-4

- 7.21 在习题 7.20 中, 求 500 名同学中有多少同学的重量满足下列要求: (a) 小于 128 磅, (b) 等于 128 磅, (c) 小于或等于 128 磅.

解 (a) 记录重量小于 128 磅的学生重量小于 127.5 磅.

$$127.5 \text{ 的标准值} = \frac{127.5 - 151}{15} = -1.57$$

如图 7-5(a)

$$\begin{aligned} \text{要求的学生比例} &= (z = -1.57 \text{ 左方的面积}) \\ &= (z = 0 \text{ 左方的面积}) - (z = -1.57 \text{ 和 } z = 0 \text{ 之间的面积}) \\ &= 0.5 - 0.4418 = 0.0582 \end{aligned}$$

则重量小于 128 磅的学生数 $= 500 \times 0.0582 = 29$.

(b) 记录重量为 128 磅的学生重量在 127.5 磅和 128.5 磅之间.

$$127.5 \text{ 的标准值} = \frac{127.5 - 151}{15} = -1.57$$

$$128.5 \text{ 的标准值} = \frac{128.5 - 151}{15} = -1.50$$

如图 7-5(b)

$$\begin{aligned} \text{要求的学生比例} &= (z = -1.57 \text{ 和 } z = -1.50 \text{ 之间的面积}) \\ &= (z = -1.57 \text{ 和 } z = 0 \text{ 之间的面积}) \\ &\quad - (z = -1.50 \text{ 和 } z = 0 \text{ 之间的面积}) \\ &= 0.4418 - 0.4332 = 0.0086 \end{aligned}$$

则重量为 128 磅的学生数 $= 500 \times 0.0086 = 4$.

(c) 记录重量小于或等于 128 磅的学生重量小于 128.5 磅.

$$128.5 \text{ 的标准值} = \frac{128.5 - 151}{15} = -1.50$$

如图 7-5(c)

$$\begin{aligned} \text{要求的学生比例} &= (z = -1.50 \text{ 左方的面积}) \\ &= (z = 0 \text{ 左方的面积}) - (z = -1.50 \text{ 和 } z = 0 \text{ 之间的面积}) \\ &= 0.5 - 0.4332 = 0.0668 \end{aligned}$$

则重量小于或等于 128 磅的学生数 = $500 \times 0.0668 = 33$.

另解(利用(a)和(b))

$$\begin{aligned} \text{重量小于或等于 128 磅的学生数} &= (\text{重量小于 128 磅的学生数}) \\ &\quad + (\text{重量等于 128 磅的学生数}) = 29 + 4 = 33. \end{aligned}$$

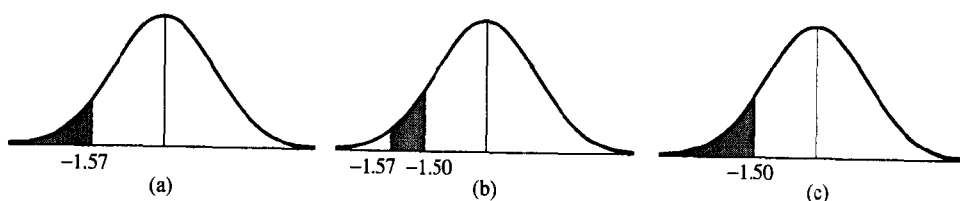


图 7-5

- 7.22** 在一次生物小测验中共有 10 个问题, 根据学生回答出的问题个数, 得分分别为 0, 1, 2, ..., 10 分. 学生得分的平均分是 6.7 分, 标准差是 1.2 分. 假设分数是服从正态分布的, 求: (a) 得 6 分的学生所占的百分比, (b) 全班得分最低的 10% 中最高的分数, (c) 全班得分最高的 10% 中最底的分数.

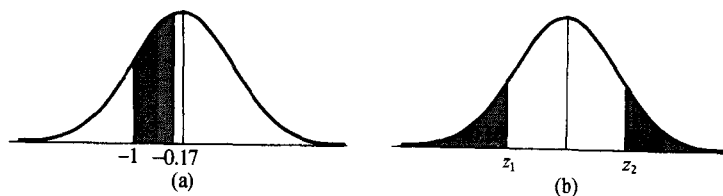


图 7-6

解 (a) 为了使离散型数据能应用正态分布, 必须将数据看成连续的, 则 6 分可看成 5.5 分到 6.5 分.

$$5.5 \text{ 的标准值} = \frac{5.5 - 6.7}{1.2} = -1.0$$

$$6.5 \text{ 的标准值} = \frac{6.5 - 6.7}{1.2} = -0.17$$

如图 7-6(a)

$$\begin{aligned} \text{要求的比例} &= (z = -1 \text{ 和 } z = -0.17 \text{ 之间的面积}) \\ &= (z = -1 \text{ 和 } z = 0 \text{ 之间的面积}) \\ &\quad - (z = -0.17 \text{ 和 } z = 0 \text{ 之间的面积}) \\ &= 0.3413 - 0.0675 = 0.2738 = 27\% \end{aligned}$$

(b) X_1 表示所求的最高分数, z_1 表示这一分数的标准值. 由图 7-6 (b), z_1 左方的面积是 10% = 0.10, 则 $(z_1 \text{ 和 } 0 \text{ 之间的面积}) = 0.40$, $z_1 = -1.28$ (很接近). 因此 $z_1 = (X_1 - 6.7)/1.2 = -1.28$, $X_1 = 5.2$, 或舍入至最近的整数 5.

(c) X_2 表示要求的最低分数, z_2 表示这一分数的标准值. 由图 7-6 (b) 及对称性, $z_2 = 1.28$. 因此 $(X_2 - 6.7)/1.2 = 1.28$, $X_2 = 8.2$, 8 是最接近的整数.

- 7.23** 一台机器生产的 200 个垫圈样品的直径的均值是 0.502 英寸, 标准差是 0.005 英寸. 生产所要求的直径的范围是 0.496 英寸到 0.508 英寸, 否则, 垫圈就是次品. 假定直径

是服从正态分布的, 求这台机器生产的垫圈的次品百分比.

解

$$0.496 \text{ 的标准值} = \frac{0.496 - 0.502}{0.005} = -1.2$$

$$0.508 \text{ 的标准值} = \frac{0.508 - 0.502}{0.005} = 1.2$$

如图 7-7

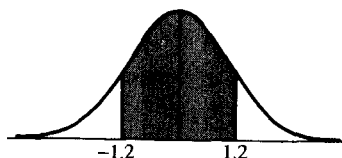


图 7-7

正品的比例 = (正态曲线下方在 $z = -1.2$ 和 $z = 1.2$ 之间的面积)
= ($z = 0$ 和 $z = 1.2$ 之间面积的两倍)
= $2 \times 0.3849 = 0.7698$ 或 77%

则次品的比例为 $100\% - 77\% = 23\%$.

注意, 如果我们认为区间 0.496 英寸到 0.508 英寸表示的实际直径是 0.4955 英寸到 0.5085 英寸, 上述的结果就要稍作修改. 而对于两位有效数字, 结果是相同的.

位有效数字, 结果是相同的.

二项分布的正态逼近

7.24 抛掷一枚均匀硬币 10 次, 分别用下列方法求出现正面 3 到 6 次的概率: (a) 二项分布, (b) 二项分布的正态逼近.

解 (a)

$$P(\text{出现正面 3 次}) = \binom{10}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^7 = \frac{15}{128}$$

$$P(\text{出现正面 4 次}) = \binom{10}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^6 = \frac{105}{512}$$

$$P(\text{出现正面 5 次}) = \binom{10}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^5 = \frac{63}{256}$$

$$P(\text{出现正面 6 次}) = \binom{10}{6} \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^4 = \frac{105}{512}$$

$$P(\text{出现正面 3 到 6 次}) = \frac{15}{128} + \frac{105}{512} + \frac{63}{256} + \frac{105}{512} \\ = \frac{99}{128} = 0.7734$$

(b) 抛掷硬币 10 次正面出现次数的概率分布如图 7-8 (a) 和 7-8 (b), 其中图 7-8 (b) 将数据看成连续的. 要求的概率就是图 7-8 (b) 中画有阴影的矩形面积之和, 它可用相应的正态曲线下方的面积来近似.

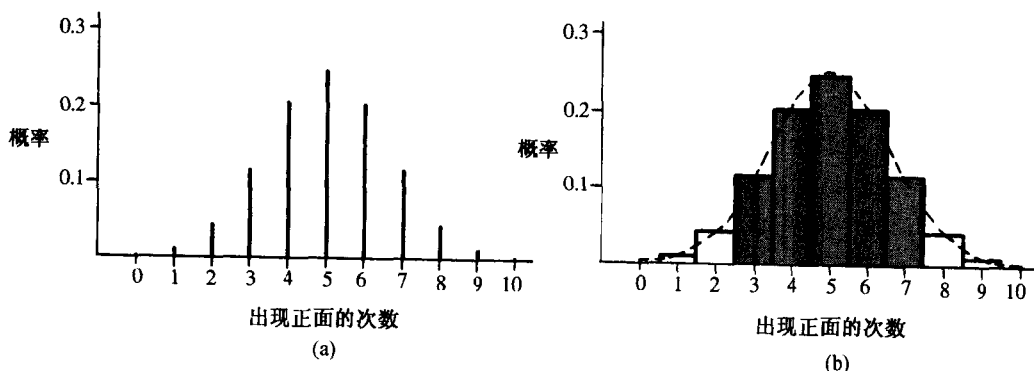


图 7-8

若将数据看成连续的, 则出现正面 3 到 6 次可看成出现正面 2.5 次到 6.5 次. 二项分布的均值和方差

$$\text{为 } \mu = Np = 10 \times \frac{1}{2} = 5 \text{ 和 } \sigma = \sqrt{Npq} = \sqrt{10 \times \frac{1}{2} \times \frac{1}{2}} = 1.58.$$

$$2.5 \text{ 的标准值} = \frac{2.5 - 5}{1.58} = -1.58$$

$$6.5 \text{ 的标准值} = \frac{6.5 - 5}{1.58} = 0.95$$

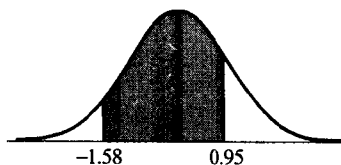


图 7-9

如图 7-9

$$\begin{aligned} \text{要求的概率} &= (z = -1.58 \text{ 和 } z = 0.95 \text{ 之间的面积}) \\ &= (z = -1.58 \text{ 和 } z = 0 \text{ 之间的面积}) \\ &\quad + (z = 0 \text{ 和 } z = 0.95 \text{ 之间的面积}) \\ &= 0.4429 + 0.3289 = 0.7718 \end{aligned}$$

这和(a)中得到的真实值 0.7734 非常接近. 当 N 更大时, 近似效果会更好.

- 7.25** 抛掷一枚均匀硬币 500 次. 求正面出现的次数和 250 相差 (a) 不多于 10, (b) 不多于 30 的概率.

$$\text{解 } \mu = Np = 500 \times \frac{1}{2} = 250 \quad \sigma = \sqrt{Npq} = \sqrt{500 \times \frac{1}{2} \times \frac{1}{2}} = 11.18$$

(a) 我们要求的是出现正面的次数在 240 和 260 之间的概率, 若将数据看作连续的, 则是在 239.5 和 260.5 之间的概率. 因为 239.5 的标准值是 $(239.5 - 250)/11.18 = -0.94$, 260.5 的标准值是 0.94, 则有

$$\begin{aligned} \text{要求的概率} &= (\text{正态曲线下方 } z = -0.94 \text{ 和 } z = 0.94 \text{ 之间的面积}) \\ &= (z = 0 \text{ 和 } z = 0.94 \text{ 之间的面积的两倍}) \\ &= 2 \times 0.3264 = 0.6528 \end{aligned}$$

(b) 我们要求的是出现正面的次数在 220 和 280 之间的概率, 若将数据看作连续的, 则是在 219.5 和 280.5 之间的概率. 因为 219.5 的标准值是 $(219.5 - 250)/11.18 = -2.73$, 280.5 的标准值是 2.73, 则有

$$\begin{aligned} \text{要求的概率} &= (z = 0 \text{ 和 } z = -2.73 \text{ 之间的面积的两倍}) \\ &= 2 \times 0.4968 = 0.9936 \end{aligned}$$

由此, 我们可以很有把握地认为出现正面的次数不会和期望 250 相差多于 30. 因此, 如果出现正面次数超过 280, 我们可以确定硬币不是均匀的 (即有负重).

- 7.26** 假设在人群中 75% 的人经常使用安全带. 随机截住 100 辆汽车, 求有 70 人或少于 70 人使用安全带的概率. 用 Minitab 对二项分布和正态分布的逼近两种方法求解.

解 如下的 Minitab 输出结果表明有 70 人或少于 70 人使用安全带的概率等于 0.1495.

MTB > cdf 70;

SUBC > binomial 100 .75.

Cumulative Distribution Function

Binomial with n=100 and p=0.750000

x	P(X ≤ x)
70.00	0.1495

用二项分布的正态逼近求解如下: 二项分布的均值为 $\mu = Np = 100 \times 0.75 = 75$, 标准差为 $\sigma = \sqrt{Npq} = \sqrt{100 \times 0.75 \times 0.25} = 4.33$. Minitab 的输出结果表明近似值等于 0.1493. 这和真实值非常接近.

MTB > cdf 70.5;

SUBC > normal mean=75 sd=4.33.

Cumulative Distribution Function

Normal with mean = 75.0000 and standard deviation = 4.33000

x	P(X ≤ x)
70.5000	0.1493

泊松分布

- 7.27 一个工厂生产的工具中有 10% 是次品, 从中随机抽取 10 个样品, 用下列方法求正好有两个次品的概率: (a) 二项分布, (b) 二项分布的泊松逼近.

解 工具是次品的概率是 $p = 0.1$.

$$(a) P(10 \text{ 个样品中有 } 2 \text{ 个次品}) = \binom{10}{2} \cdot 0.1^2 \cdot 0.9^8 = 0.1937 \text{ 或 } 0.19$$

$$(b) \text{ 由 } \lambda = np = 10 \times 0.1 = 1 \text{ 和 } e = 2.718$$

$$P(10 \text{ 个样品中有 } 2 \text{ 个次品}) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1^2 \cdot e^{-1}}{2!} = \frac{e^{-1}}{2} = \frac{1}{2e} = 0.1839 \text{ 或 } 0.18$$

一般说来, 如果 $p \leq 0.1$ 且 $\lambda = np \leq 5$, 近似效果会更好.

- 7.28 如果一个人注射一种药剂产生不良反应的概率是 0.001, 在 2000 人中求 (a) 正好有 3 人, (b) 多于 2 人产生不良反应的概率. 用 Minitab 对泊松分布和二项分布两种方法求解.

解 (a) 如下的 Minitab 结果先给出了正好有 3 人产生不良反应的二项分布概率. 由 $\lambda = np = 2000 \times 0.001 = 2$, 接着又给出了泊松分布概率. 可见泊松逼近和二项分布概率非常接近.

MTB > pdf 3;

SUBC > binomial 2000 .001.

Probability Density Function

Binomial with n = 2000 and p = 0.001

x	P(X = x)
3.0	0.1805

MTB > pdf 3;

SUBC > poisson 2.

Probability Density Function

Poisson with mu = 2

x	P(X = x)
3.00	0.1804

(b) 多于 2 人产生不良反应的概率为 $1 - P(X \leq 2)$. 如下的 Minitab 结果说明用二项分布和泊松分布求出的 $X \leq 2$ 的概率都是 0.6767. 因此, 多于 2 人产生不良反应的概率是 $1 - 0.6767 = 0.3233$.

MTB > cdf 2;

SUBC > binomial 2000 .001.

Cumulative Distribution Function

Binomial with n = 2000 and p = 0.001

x	P(X ≤ x)
2.0	0.6767

MTB > cdf 2;

SUBC > poisson 2.

Cumulative Distribution Function

Poisson with mu = 2

x	P(X ≤ x)
2.00	0.6767

- 7.29 一个泊松分布为

$$p(X) = \frac{0.72^x e^{-0.72}}{X!}$$

求 (a) $p(0)$, (b) $p(1)$, (c) $p(2)$, (d) $p(3)$.

解 (a) 由附录 VII $p(0) = \frac{0.72^0 e^{-0.72}}{0!} = \frac{1 \cdot e^{-0.72}}{1} = e^{-0.72} = 0.4868$

$$\begin{aligned}
 (b) \quad p(1) &= \frac{0.72^1 e^{-0.72}}{1!} = 0.72 \cdot e^{-0.72} = 0.72 \times 0.4868 = 0.3505 \\
 (c) \quad p(2) &= \frac{0.72^2 e^{-0.72}}{2!} = \frac{0.5184 \cdot e^{-0.72}}{2} = 0.2592 \times 0.4868 = 0.1262 \\
 \text{另解} \quad p(2) &= \frac{0.72}{2} p(1) = 0.36 \times 0.3505 = 0.1262 \\
 (d) \quad p(3) &= \frac{0.72^3 e^{-0.72}}{3!} = \frac{0.72}{3} p(2) = 0.24 \times 0.1262 = 0.0303
 \end{aligned}$$

多项分布

7.30 一个盒子里有 5 个红球, 4 个白球和 3 个蓝球. 从盒子里随机取出一个球并记下它的颜色, 然后再放回盒子. 按这种方法从中取出 6 个球, 求其中有 3 个红球, 2 个白球和 1 个蓝球的概率.

解 $P(\text{任一次取出的是红球}) = \frac{5}{12}, P(\text{任一次取出的是白球}) = \frac{4}{12},$
 $P(\text{任一次取出的是蓝球}) = \frac{3}{12}.$ 则

$$P(3 \text{ 个红球}, 2 \text{ 个白球和 } 1 \text{ 个蓝球}) = \frac{6!}{3!2!1!} \left(\frac{5}{12}\right)^3 \left(\frac{4}{12}\right)^2 \left(\frac{3}{12}\right)^1 = \frac{625}{5184}$$

用数据拟合理论分布

7.31 拟合一个二项分布使之适合习题 2.17 中的数据.

解 我们已知 $P(\text{抛掷 } 5 \text{ 个硬币 } 1 \text{ 次有 } X \text{ 个出现正面}) = p(X) = \binom{5}{X} p^X q^{5-X},$ 其中 p 和 q 分别表示抛掷一个硬币一次正面和反面出现的概率. 由习题 7.11 (a), 出现正面的平均数是 $\mu = Np = 5p.$ 对于实际(或观测)的频数分布, 出现正面的平均数是

$$\begin{aligned}
 \frac{\sum fX}{\sum f} &= \frac{38 \times 0 + 144 \times 1 + 342 \times 2 + 287 \times 3 + 164 \times 4 + 25 \times 5}{1000} = \frac{2470}{1000} \\
 &= 2.47
 \end{aligned}$$

使理论上的均值和实际中的均值相等, $5p = 2.47,$ 或 $p = 0.494.$ 则拟合的二项分布为 $p(X) = \binom{5}{X} 0.494^X 0.506^{5-X}.$

表 7.4 列出了这些概率以及期望中(理论上)的和实际中的频数. 可见拟合的效果挺好. 拟合的优度问题将在习题 12.12 中讨论.

表 7.4

出现正面的次数(X)	$P(\text{出现正面 } X \text{ 次})$	期望的频数	观测的频数
0	0.0332	33.2 或 33	38
1	0.1619	161.9 或 162	144
2	0.3162	316.2 或 316	342
3	0.3087	308.7 或 309	287
4	0.1507	150.7 或 151	164
5	0.0294	29.4 或 29	25

7.32 用概率图纸判断表 2.1 中的频数分布能否用正态分布较好地近似.

解 首先将给定的频数分布转变为一个累积频率分布, 如表 7.5 所示. 然后在特定的概率图纸上画出用百分数表示的累积频率分布对上界的散点图, 如图 7-10. 图中所画点的共线程度决定了给定的分布能否用一个正态分布很好地拟合. 可见, 存在一个正态分布能够很好地拟合数据(见习题 7.33).

表 7.5

身高(英寸)	累积频率(%)
小于 62.5	5.0
小于 65.5	23.0(c_1)
小于 68.5	65.0
小于 71.5	92.0
小于 74.5	100.0

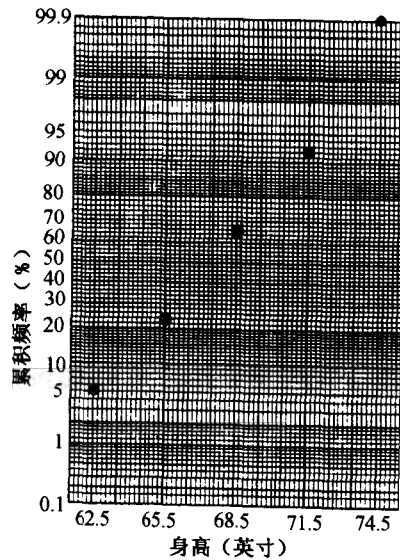


图 7-10

7.33 拟合一个正态曲线使之适合表 2.1 中的数据.

表 7.6

身高(英寸)	组界(X)	组界的 z 值	正态曲线下 0 和 z 之间的面积	每组的面积	期望的频数	观测的频数
60~62	59.5	-2.72	0.4967	0.0413	4.13 或 4	5
63~65	62.5	-1.70	0.4554	0.2068	20.68 或 21	18
66~68	65.5	-0.67	0.2486	0.3892	38.92 或 39	42
69~71	68.5	0.36	0.1406			
72~74	71.5	1.39	0.4177	0.0743	7.43 或 7	8
	74.5	2.41	0.4920			

$$\bar{X} = 67.45 \text{ 英寸} \quad s = 2.92 \text{ 英寸}$$

解 可按表 7.6 进行分组. 用 $z = (X - \bar{X})/s$ 计算每一组的组界, 其中均值 \bar{X} 和标准差 s 已分别在习题 3.22 和 4.17 中求得.

在表 7.6 中的第 4 列, 正态曲线下 0 和 z 之间的面积已从附录 II 中查到. 由此, 可求出正态曲线下在连续的 z 值之间的面积, 如第 5 列所示. 得到这些面积的方法是: 当连续的 z 值同号时, 将相应的第 4 列中的连续的面积相减; 异号时则相加(表中只出现一次这样的情况). 由图形可以很容易地看出这样做的理由.

将第 5 列的值(表示频率)乘以总频数 N (此时 $N = 100$)得到期望的频数, 如第 6 列所示. 可见它

们都和第 7 列中的真实(或观测)的频数很吻合.

如有需要,可用 Sheppard 的方法修正标准差(见习题 4.21(a)).

这一分布的拟合优度将在习题 12.13 中讨论.

- 7.34 表 7.7 表示的是一座城市在 50 天里每天发生 X 起交通事故的天数. 拟合一个泊松分布使之适合这些数据.

表 7.7

交通事故的次数(X)	天数(f)
0	21
1	18
2	7
3	3
4	1
	和 50

解 事故的平均数是

$$\lambda = \frac{\sum fX}{\sum f} = \frac{21 \times 0 + 18 \times 1 + 7 \times 2 + 3 \times 3 + 1 \times 4}{50} = \frac{45}{50} = 0.90$$

则根据泊松分布

$$P(X \text{ 起事故}) = \frac{0.90^X e^{-0.90}}{X!}$$

表 7.8 列出了由这个泊松分布算出的发生 0, 1, 2, 3 和 4 起事故的概率以及期望中或理论上发生 X 起事故的天数(分别用概率乘以 50). 为了便于比较, 第 4 列中重复列出了表 7.7 中的实际天数.

可见泊松分布是所给数据的较好拟合.

对于一个真正的泊松分布, 方差 $\sigma^2 = \lambda$. 计算给定的频数分布的方差得到 0.97. 将此值与 λ 的值 0.90 比较, 进一步说明了泊松分布是样本数据的较好的拟合.

表 7.8

交通事故的次数(X)	P (发生 X 起交通事故)	期望的天数	实际天数
0	0.4066	20.33 或 20	21
1	0.3659	18.30 或 18	18
2	0.1647	8.24 或 8	7
3	0.0494	2.47 或 2	3
4	0.0111	0.56 或 1	1

补充习题

二项分布

- 7.35 计算(a) $7!$, (b) $\frac{10!}{6! 4!}$, (c) $\binom{9}{5}$, (d) $\binom{11}{8}$, (e) $\binom{6}{1}$.
- 7.36 将下列各式展开 (a) $(q+p)^7$, (b) $(q+p)^{10}$.
- 7.37 抛掷一枚均匀硬币 6 次, 求分别有(a) 0, (b) 1, (c) 2, (d) 3, (e) 4, (f) 5, (g) 6 次出现正面的概率.
- 7.38 抛掷一枚均匀硬币 6 次, 求分别有(a) 不少于 2 次, (b) 少于 4 次出现正面的概率.
- 7.39 如果用 X 表示抛掷 4 枚均匀硬币 1 次出现正面的硬币个数, 求(a) $P(X=3)$, (b) $P(X<2)$, (c) $P(X \leq 2)$, (d) $P(1 < X \leq 3)$.
- 7.40 在拥有 5 个孩子的 800 个家庭中, 预计有多少个家庭有(a) 3 个男孩, (b) 5 个女孩, (c) 2 个或 3 个男孩? 假定男孩和女孩出生的概率是相同的.

- 7.41 抛掷一对均匀骰子两次,求有(a)一次,(b)两次点数和为11的概率.
- 7.42 抛掷一对均匀骰子三次,求恰好有一次点数和为9的概率.
- 7.43 在有10题的是非题测验中,至少答对6题的概率是多少?
- 7.44 一位保险经纪人向5个人卖出了保险,他们的年龄相同且身体状况良好.根据保险表,这个年龄的人再活30年的概率是 $\frac{2}{3}$.求30年后,(a)5个人,(b)至少3个人,(c)只有2个人,(d)至少1个人在世的概率.
- 7.45 计算二项分布的(a)均值,(b)标准差,(c)矩偏度系数,(d)矩峰度系数.其中 $p=0.7$, $N=60$.并解释所得的结果.
- 7.46 证明:如果一个 $N=100$ 的二项分布是对称的,则它的矩峰度系数是2.98.
- 7.47 对于二项分布,计算(a) $\sum (X-\mu)^3 p(X)$, (b) $\sum (X-\mu)^4 p(X)$.
- 7.48 用本章开始处提到的公式(1)和公式(2)计算矩偏度系数和矩峰度系数.

正态分布

- 7.49 一次统计学测验的,均分是78,标准差是10.
(a)求93分和62分分别对应的标准分数.
(b)求标准分为-0.6和1.2所对应的分数.
- 7.50 在一次测验中,70分和88分对应的标准分数分别是-0.6和1.4,求(a)均分,(b)标准差.
- 7.51 求正态曲线下方在(a) $z=-1.20$ 和 $z=2.40$ 之间,(b) $z=1.23$ 和 $z=1.87$ 之间,(c) $z=-2.35$ 和 $z=-0.50$ 之间的面积.
- 7.52 求正态曲线下方(a)在 $z=-1.78$ 的左方,(b)在 $z=0.56$ 的左方,(c)在 $z=-1.45$ 的右方,(d)对应于 $z \geq 2.16$, (e)对应于 $-0.80 \leq z \leq 1.53$, (f)在 $z=-2.52$ 的左方, $z=1.83$ 的右方的面积.
- 7.53 如果 z 服从均值为0,方差为1的正态分布,求(a) $P(z \geq -1.64)$, (b) $P(-1.96 \leq z \leq 1.96)$, (c) $P(|z| \geq 1)$.
- 7.54 求 z 值,使得(a) z 右方的面积是0.2266, (b) z 左方的面积是0.0314, (c) -0.23 和 z 之间的面积是0.5722, (d) 1.15 和 z 之间的面积是0.0730, (e) $-z$ 和 z 之间的面积是0.9000 其中 z 服从均值为0,方差为1的正态分布.
- 7.55 求 z_1 使得 $P(z \geq z_1) = 0.84$, 其中 z 服从均值为0,方差为1的正态分布.
- 7.56 求正态曲线在(a) $z=2.25$, (b) $z=-0.32$, (c) $z=-1.18$ 处的纵坐标.
- 7.57 如果300名学生的身高服从均值为68.0英寸,标准差为3.0英寸的正态分布.那么有多少名学生的身高满足下列条件?假定记录测量结果时保留到最近的整数.(a)大于72英寸,(b)小于或等于64英寸,(c)65英寸和71英寸之间,(d)等于68英寸.
- 7.58 如果轴承滚珠的直径服从均值为0.6140英寸,标准差为0.0025英寸的正态分布.求轴承滚珠的直径满足下列条件的百分比:(a) 0.610英寸和0.618英寸之间,(b)大于0.617英寸,(c)小于0.608英寸,(d)等于0.615英寸.
- 7.59 一次期终考试的平均分是72,标准差是9.得分在前10%的学生得A.一名学生要想得A,最低要得多少分?
- 7.60 假设一组测量数据服从正态分布,则满足下列条件的数据所占的百分比是多少?(a)和均值的差大于标准差的一半,(b)和均值的差小于标准差的四分之三.
- 7.61 假设一组测量数据服从均值为 \bar{X} ,标准差为 s 的正态分布.则满足下列条件的数据所占的百分比是多少?(a)在 $\bar{X} \pm 2s$ 之内,(b)在 $\bar{X} \pm 1.2s$ 之外,(c)大于 $\bar{X} - 1.5s$.
- 7.62 在习题7.61中,求常数 a 使得百分比满足下列条件:(a)在 $\bar{X} \pm as$ 之内的占75%,(b)小于 $\bar{X} - as$ 的占22%.

二项分布的正态逼近

- 7.63 抛掷一枚均匀硬币200次,求下列事件发生的概率:(a)80至120次出现正面,(b)少于90次出现正面,(c)少于85次或多于115次出现正面,(d)正好100次出现正面.
- 7.64 在一次是非题测验中,求一名学生猜对题的情况满足下列条件的概率:(a)20题中能猜对12题或12题以上,(b)40题中能猜对24题或24题以上.

- 7.65 一台机器生产的螺栓中有 10% 是次品, 从中随机抽取 400 个样品, 求次品数量满足下列条件的概率:
(a) 至多 30 个, (b) 30 至 50 个, (c) 35 至 45 个, (d) 不少于 55 个.
- 7.66 抛掷一对均匀骰子 100 次, 求 25 次以上得到点数和为 7 的概率.

泊松分布

- 7.67 假设一家公司生产的灯泡中有 3% 是次品, 求 100 个样品中次品数满足下列条件的概率: (a) 0 个, (b) 1 个, (c) 2 个, (d) 3 个, (e) 4 个, (f) 5 个.
- 7.68 在习题 7.67 中, 求次品数满足下列条件的概率: (a) 多于 5 个, (b) 1 至 3 个, (c) 不多于 2 个.
- 7.69 一个袋子里装有 1 个红球和 7 个白球, 从中有放回地抽球 8 次, 每次 1 个. 分别用下列两种方法求从中抽取 8 个球恰好有 3 个红球的概率: (a) 二项分布, (b) 二项分布的泊松逼近.
- 7.70 根据美国健康教育福利部门的国家生死统计办公室的数据显示, 在美国每年每 100 000 人中发生 3.0 起溺水事故. 求在一个拥有 200 000 人口的城市里每年发生的溺水事故的次数满足下列条件的概率: (a) 0 次, (b) 2 次, (c) 6 次, (d) 8 次, (e) 4 次和 8 次之间, (f) 少于 3 次.
- 7.71 在下午 2 点和 4 点之间, 某公司总机每分钟接到 2.5 个电话. 求在某一分钟内电话个数满足下列条件的概率: (a) 0 个, (b) 1 个, (c) 2 个, (d) 3 个, (e) 不多于 4 个, (f) 多于 6 个.

多项分布

- 7.72 抛掷一个均匀骰子 6 次, 求下列事件发生的概率: (a) 出现 1 个 1, 2 个 2 和 3 个 3, (b) 每一面出现一次.
- 7.73 一个盒子里有大量的红球、白球、蓝球和黄球, 所占的比例为 4:3:2:1. 从中抽取 10 个球, 求抽到下列球的概率: (a) 4 个红球, 3 个白球, 2 个蓝球和 1 个黄球, (b) 8 个红球和 2 个黄球.
- 7.74 抛掷一个均匀骰子 4 次, 求一次也得不到 1, 2 或 3 的概率.

用数据拟合理论分布

- 7.75 用表 7.9 中的数据拟合二项分布.

表 7.9

X	0	1	2	3	4
f	30	62	46	10	2

- 7.76 用概率图纸判断习题 3.59 中的数据的频率分布能否用正态分布很好地近似.
- 7.77 用习题 3.59 中的数据拟合正态分布.
- 7.78 用习题 3.61 中的数据拟合正态分布.
- 7.79 用习题 7.75 中的数据拟合泊松分布, 并和二项分布的拟合相比较.
- 7.80 表 7.10 列出了 20 年间(1875 至 1894)10 支普鲁士军队中每年被马踢死的人数. 用这些数据拟合泊松分布.

表 7.10

X	0	1	2	3	4
f	109	65	22	3	1

第八章 初等抽样理论

抽样理论

抽样理论研究的是总体和从总体中抽取的样本之间的关系,在许多方面都有重要的应用价值.例如,可通过样本的某些量(如样本均值和方差)来估计总体的相应量(如总体均值和方差),即估计**总体参数**或简称**参数**,这就是**抽样统计**,简称**统计**.有关估计的问题将在第九章讨论.

抽样理论还可用以判断两样本之间的差异性是由于偶然的偏差还是它们之间真的有差异.例如在检验某种新药是否有效或判断某种生产线是否优于其他生产线时,常会遇到类似问题.回答这些问题就要用到所谓的**显著性检验**和**假设检验**,这在**决策论**中是非常重要的,将在第十章予以讨论.

一般说来,利用从总体中抽取的样本对总体作出推断以及用概率论的理论对这种推断的准确性作出提示,这些就称为**统计推断**.

随机样本和随机数

为了保证抽样理论和统计推断的结论能够成立,选择的样本必须是总体的**代表**.由此引出的对抽样方法以及相关问题的研究称为**试验设计**.

获得有代表性的样本的方法之一是**随机抽样**,即总体中的个体被抽到的机会是均等的.可以给总体中的每个个体编一个号码,将这些数字写在小纸片上,放入罐子里,然后再从中抽取.要注意的是在每次抽取之前都要搅拌充分.另一个可供选择的方法是用为此专门设计的**随机数表**(见附录Ⅸ).可参见习题 8.6.

有放回和无放回抽样

从罐子里抽取一个数后,在抽取第二个数之前我们可以选择将该数字放回或不放回.在第一种情况下,一个数字可以一次又一次重复出现,而在第二种情况下,每个数字只能出现一次.总体中的每个个体可以不止一次地被选中的抽样叫做**有放回抽样**,如果每个个体被选中的次数不多于一次,则叫**无放回抽样**.

总体可以是有限的,也可以是无限的.比如说,如我们从一个装有 100 个球的罐子里无放回地连续抽取 10 个球,那就是从一个有限的总体中抽样;如果我们抛掷一枚硬币 50 次,记录正面出现的次数,我们就是从一个无限的总体中抽样.

从理论上来说,对于有放回抽样,由于无论从总体中抽取了多少个样本,都不会使总体变小,有限的总体也可看成是无限的.出于许多实际的考虑,当总体很大时,从有限的总体中抽样也可看成是从无限的总体中抽样.

抽样分布

考虑从一个给定的总体中抽取(不论是否有放回)容量(或大小)为 N 的所有可能的样本.对于每一个样本,我们可计算出某个统计量(如样本均值或标准差)的值,不同的样本得到的该统计量的值是不一样的.用这样的方法我们能得到这个统计量的分布,称之为**抽样分布**.

例如,如果特指的统计量是样本均值,则此分布称为**均值的抽样分布**.类似地,我们可以得到标准差、方差、中位数、比例的抽样分布等.

对于每个统计量的抽样分布,可计算出它的均值和标准差等.我们称之为该统计量抽样分布的均值和标准差等,或简称为该统计量的均值和标准差等.

均值的抽样分布

假设从一个容量为 N_p 的有限总体中无放回地抽取了所有的容量为 N 的样本, $N_p > N$. 分别用 $\mu_{\bar{X}}$ 和 $\sigma_{\bar{X}}$ 记均值的抽样分布的均值和标准差, 用 μ 和 σ 记总体的均值和标准差, 则

$$\mu_{\bar{X}} = \mu \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} \quad (1)$$

如果总体是无限的或抽取是有放回的, 上述结果变为

$$\mu_{\bar{X}} = \mu \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \quad (2)$$

当 N 较大时 ($N \geq 30$), 无论总体如何 (只要总体的均值和方差有限且总体的容量至少是样本容量的两倍), 均值的抽样分布都近似于一个均值为 $\mu_{\bar{X}}$, 标准差为 $\sigma_{\bar{X}}$ 的正态分布. 若总体是无限的, 这一结论就是现代概率论中的**中心极限定理**的一个特例, 它说明 N 越大, 这种近似效果越好. 我们有时也说这个抽样分布是**近似正态的**.

如果总体服从正态分布, 即使 N 较小 (如 $N < 30$), 均值的抽样分布也是正态的.

比例的抽样分布

假设总体是无限的且某事件发生 (称为成功) 的概率是 p , 而该事件不发生 (称为失败) 的概率是 $q = 1 - p$. 例如, 总体是抛掷一枚均匀硬币任意次的结果, 其中正面出现的概率是 $p = \frac{1}{2}$. 考虑从中抽取的所有容量为 N 的样本, 并对每个样本求该事件发生的比例 P . 例如, P 就是抛掷一枚均匀硬币 N 次正面出现的比例. P 的抽样分布称为**比例的抽样分布**, 它的均值 μ_P 和标准差 σ_P 分别为

$$\mu_P = p \quad \sigma_P = \sqrt{\frac{pq}{N}} = \sqrt{\frac{p(1-p)}{N}} \quad (3)$$

将 $\mu = p$ 和 $\sigma = \sqrt{pq}$ 代入 (2) 式也可得上式. 当 N 很大时 ($N \geq 30$), 这个抽样分布非常接近于正态分布. 注意, 总体服从**二项分布**.

当总体有限且抽样有放回时, (3) 式也成立. 而当总体有限且抽样无放回时, (3) 式不再成立, 成立的是将 $\mu = p$ 和 $\sigma = \sqrt{pq}$ 代入 (1) 式所得的式子.

显然, 用 N 除二项分布的均值和标准差 (Np 和 \sqrt{Npq}) 即得 (3) 式 (见第七章).

差与和的抽样分布

假设有两个给定的总体, 对于从第一个总体中抽取的每一个容量为 N_1 的样本, 可计算统计量 S_1 . 进而可求出 S_1 的抽样分布, 它的均值和标准差分别记为 μ_{S_1} 和 σ_{S_1} . 同样地, 对于从第二个总体中抽取的每一个容量为 N_2 的样本, 可计算统计量 S_2 . 进而可求出 S_2 的抽样分布, 它的均值和标准差分别记为 μ_{S_2} 和 σ_{S_2} . 结合从两个总体中抽取的样本, 我们可得到差值 $S_1 - S_2$ 的分布, 称为**统计量之差的抽样分布**. 这一分布的均值和标准差分别记为 $\mu_{S_1 - S_2}$ 和 $\sigma_{S_1 - S_2}$, 如果所抽取的样本互不依赖 (即样本是独立的), 则有

$$\mu_{S_1 - S_2} = \mu_{S_1} - \mu_{S_2} \quad \sigma_{S_1 - S_2} = \sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2} \quad (4)$$

如果 S_1 和 S_2 指的是两个总体的样本均值, 分别用 \bar{X}_1 和 \bar{X}_2 表示, 那么对于均值和标准差分别为 (μ_1, σ_1) 和 (μ_2, σ_2) 的两个无限的总体, 均值之差的抽样分布是一定的, 利用 (2) 式可得

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2 \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \quad (5)$$

如果总体有限但抽取是有放回的,结论仍然成立.如果总体有限但抽取是无放回的,可由(1)式得到类似的结论.

对于两个参数分别为 (p_1, q_1) 和 (p_2, q_2) 的服从二项分布的总体,关于比例之差的抽样分布也能得到相应的结论.此时, S_1 和 S_2 相当于成功的比例 P_1 和 P_2 , 由(4)式可得

$$\mu_{P_1-P_2} = \mu_{P_1} - \mu_{P_2} = p_1 - p_2, \quad \sigma_{P_1-P_2} = \sqrt{\sigma_{P_1}^2 + \sigma_{P_2}^2} = \sqrt{\frac{p_1 q_1}{N_1} + \frac{p_2 q_2}{N_2}} \quad (6)$$

若 N_1 和 N_2 较大($N_1, N_2 \geq 30$), 均值之差或比例之差的抽样分布都接近于正态分布.

有时也要用到统计量之和的抽样分布. 如果样本是独立的, 这一分布的均值和标准差为

$$\mu_{S_1+S_2} = \mu_{S_1} + \mu_{S_2}, \quad \sigma_{S_1+S_2} = \sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2} \quad (7)$$

标准误差

一个统计量的抽样分布的标准差常称为该统计量的**标准误差**. 表 8.1 列出了从无限总体(或很大总体)中随机抽样或从有限总体中有放回抽样所得的不同统计量的抽样分布的标准误差. 表中也注明了结论成立所需的条件以及其他一些重要的结论.

μ, σ, p, μ_r 和 \bar{X}, s, P, m_r 分别表示总体和样本的均值、标准差、比例及均值的 r 阶矩.

要注意的是, 当样本容量足够大时, 抽样分布是正态或接近于正态的, 与之相关的方法称为**大样本方法**. 当 $N < 30$ 时, 样本容量较小, 与之相关的理论称为**精确抽样理论**, 这将在第十一章讨论.

若样本的某些参数如 σ, p 或 μ_r 未知, 而样本容量足够大, 则可用与之相对应的样本统计量 s (或 $s = \sqrt{N/(N-1)}s$), P 和 m_r 对其作出较好的估计.

表 8.1 抽样分布的标准误差

抽样分布	标准误差	注释
均值	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$	对于小样本和大样本结论均成立. 当 $N \geq 30$ 时, 即使总体是非正态的, 均值的抽样分布也是非常接近于正态分布. 在所有情况下, $\mu_{\bar{X}} = \mu$, 即等于总体均值
比例	$\sigma_P = \sqrt{\frac{p(1-p)}{N}} = \sqrt{\frac{pq}{N}}$	对于均值的注释也适用于此 在所有情况下, $\mu_P = p$
标准差	$(1) \sigma_s = \frac{\sigma}{\sqrt{2N}}$ $(2) \sigma_s = \sqrt{\frac{\mu_4 - \mu_2^2}{4N\mu_2}}$	当 $N \geq 100$ 时, s 的抽样分布是非常接近于正态的. 若总体是正态的(或近似于正态), σ_s 由(1)式给出. 若总体是非正态的, σ_s 由(2)式给出. 注意: 当总体服从正态分布时, (2)式中的 $\mu_2 = \sigma^2, \mu_4 = 3\sigma^4$, (2)式简化为(1)式. 当 $N \geq 100$ 时, $\mu_s = \sigma$ 近似成立.
中位数	$\sigma_{med} = \sigma \sqrt{\frac{\pi}{2N}} = \frac{1.2533\sigma}{\sqrt{N}}$	当 $N \geq 30$ 时, 中位数的抽样分布是非常接近于正态的. 只有当总体服从正态(或近似于正态)分布时, 所给的结论才能成立. $\mu_{med} = \mu$
第 1 个四分位数和 第 3 个四分位数	$\sigma_{Q1} = \sigma_{Q3} = \frac{1.3626\sigma}{\sqrt{N}}$	对于中位数的注释也适用于此. μ_{Q1} 和 μ_{Q3} 分别与总体的第 1 个四分位数和第 3 个四分位数近似相等. 注意: $\sigma_{Q2} = \sigma_{med}$

续表

抽样分布	标准误差	注释
十分位数	$\sigma_{D1} = \sigma_{D9} = \frac{1.7094\sigma}{\sqrt{N}}$	对于中位数的注释也适用于此。 $\mu_{D1}, \mu_{D2}, \dots$ 分别与总体的第 1 个, 第 2 个, \dots 十分位数近似相等。 注意: $\sigma_{D5} = \sigma_{\text{med}}$
	$\sigma_{D2} = \sigma_{D8} = \frac{1.4288\sigma}{\sqrt{N}}$	
	$\sigma_{D3} = \sigma_{D7} = \frac{1.3180\sigma}{\sqrt{N}}$	
	$\sigma_{D4} = \sigma_{D6} = \frac{1.2680\sigma}{\sqrt{N}}$	
半内四分距	$\sigma_Q = \frac{0.7867\sigma}{\sqrt{N}}$	对于中位数的注释也适用于此。 μ_Q 和总体的半内四分距近似相等
方差	(1) $\sigma_S^2 = \sigma^2 \sqrt{\frac{2}{N}}$	对于标准差的注释也适用于此. 注意: 当总体服从正态分布时, (2) 式可化为 (1) 式。 $\mu_S^2 = \sigma^2(N-1)/N$, 当 N 较大时, 它和 σ^2 很接近.
	(2) $\sigma_S^2 = \sqrt{\frac{\mu_4 - \frac{N-3}{N-1}\mu^2}{N}}$	
变异系数	$\sigma_v = \frac{v}{\sqrt{2N}} \sqrt{1+2v^2}$	$v = \sigma/\mu$ 是总体的变异系数. 只有当 $N \geq 100$ 和总体为正态(或近似正态)时, 结论才成立.

习题及解答

均值的抽样分布

- 8.1 一个总体包含有 5 个数, 即 2, 3, 6, 8 和 11. 考察从总体中有放回抽取得到的所有容量为 2 的样本. 求 (a) 总体的均值, (b) 总体的标准差, (c) 样本均值抽样分布的均值, (d) 样本均值抽样分布的标准差(即样本均值的标准误差).

解 (a) $\mu = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6.0$

(b)

$$\begin{aligned}\sigma^2 &= \frac{(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2}{5} \\ &= \frac{16+9+0+4+25}{5} = 10.8\end{aligned}$$

则 $\sigma = 3.29$.

(c) 从总体中有放回抽取得到的容量为 2 的样本共有 25 个(因为第一次抽到的 5 个数中的任一个可和第二次抽到的 5 个数中的任一个相结合考虑). 它们是

(2, 2)	(2, 3)	(2, 6)	(2, 8)	(2, 11)
(3, 2)	(3, 3)	(3, 6)	(3, 8)	(3, 11)
(6, 2)	(6, 3)	(6, 6)	(6, 8)	(6, 11)
(8, 2)	(8, 3)	(8, 6)	(8, 8)	(8, 11)
(11, 2)	(11, 3)	(11, 6)	(11, 8)	(11, 11)

相应的样本均值是

2.0	2.5	4.0	5.0	6.5
2.5	3.0	4.5	5.5	7.0
4.0	4.5	6.0	7.0	8.5
5.0	5.5	7.0	8.0	9.5
6.5	7.0	8.5	9.5	11.0

(8)

样本均值抽样分布的均值是

$$\mu_{\bar{X}} = \frac{(8) \text{ 式中所有样本均值的和}}{25} = \frac{150}{25} = 6.0$$

说明 $\mu_{\bar{X}} = \mu$.

(d) 要求样本均值抽样分布的方差 $\sigma_{\bar{X}}^2$, 先将(8)式中的每一个数字减去均值 6, 将所得的结果平方, 再把得到的 25 个数字相加所得的和除以 25. 最终的结果是 $\sigma_{\bar{X}}^2 = 135/25 = 5.40$, 则 $\sigma_{\bar{X}} = \sqrt{5.40} = 2.32$. 这一结果说明: 对于有限总体中的有放回抽样(或无限总体), $\sigma_{\bar{X}}^2 = \sigma^2/N$. 等式右边得 $10.8/2 = 5.40$, 和上面求得的值很吻合.

8.2 如果抽样不是有放回的, 求解习题 8.1.

解 如习题 8.1 中的(a)和(b), $\mu = 6$ 且 $\sigma = 3.29$.

(c) 从总体中无放回(即先抽出一个数, 抽到的第二个数和第一个不同)地抽取所有容量为 2 的样本共有 $\binom{5}{2} = 10$ 个: (2, 3), (2, 6), (2, 8), (2, 11), (3, 6), (3, 8), (3, 11), (6, 8), (6, 11) 和 (8, 11). 例如, 抽到 (2, 3) 和抽到 (3, 2) 可认为是一样的.

相应的样本均值是 2.5, 4.0, 5.0, 6.5, 4.5, 5.5, 7.0, 7.0, 8.5 和 9.5. 样本均值抽样分布的均值是

$$\mu_{\bar{X}} = \frac{2.5 + 4.0 + 5.0 + 6.5 + 4.5 + 5.5 + 7.0 + 7.0 + 8.5 + 9.5}{10} = 6.0$$

说明 $\mu_{\bar{X}} = \mu$.

(d) 样本均值抽样分布的方差是

$$\sigma_{\bar{X}}^2 = \frac{(2.5 - 6.0)^2 + (4.0 - 6.0)^2 + (5.0 - 6.0)^2 + \cdots + (9.5 - 6.0)^2}{10} = 4.05$$

则 $\sigma_{\bar{X}} = 2.01$. 而由(1)式得到

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{N} \left(\frac{N_p - N}{N_p - 1} \right) = \frac{10.8}{2} \times \frac{5 - 2}{5 - 1} = 4.05$$

两者完全吻合

8.3 假设一所大学的 3000 名男学生的身高服从均值为 68.0 英寸, 标准差为 3.0 英寸的正态分布. 如果有 80 组样本, 每组有 25 名学生. 若(a) 抽取是有放回的, (b) 抽取是无放回的, 样本均值抽样分布的期望的均值和标准差是多少?

解 从理论上来说, 能够从 3000 名学生中有放回和无放回地抽取容量为 25 的样本各有 3000^{25}

个和 $\binom{3000}{25}$ 个, 远大于 80. 因此, 我们不能得到样本均值的真正的抽样分布, 而只能得到经验的抽样分布. 不仅如此, 由于样本的数量很大, 这两种抽样分布应该很接近. 因此可按(1), (2)式来计算样本均值经验抽样分布的期望的均值和标准差.

$$(a) \mu_{\bar{X}} = \mu = 68.0 \text{ 英寸且 } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{3}{\sqrt{25}} = 0.6 \text{ 英寸}$$

$$(b) \mu_{\bar{X}} = 68.0 \text{ 英寸且 } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} = \frac{3}{\sqrt{25}} \sqrt{\frac{3000 - 25}{3000 - 1}} \text{ 英寸}$$

这个值只比 0.6 稍微小一点, 因此对于本问题来说无放回抽样和有放回抽样可看作同样的.

所以, 我们可预计样本均值的经验抽样分布近似于均值为 68.0, 标准差为 0.6 的正态分布.

8.4 在习题 8.3 中, 预计能找到多少个样本, 使得它的均值(a) 在 66.8 英寸和 68.3 英寸之间, (b) 小于 66.4 英寸?

解 此时, 一个样本均值的标准值为

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - 68.0}{0.6}$$

(a)

$$66.8 \text{ 的标准值} = \frac{66.8 - 68.0}{0.6} = -2.0$$

$$68.3 \text{ 的标准值} = \frac{68.3 - 68.0}{0.6} = 0.5$$

如图 8-1(a)所示

$$\begin{aligned}
 & \text{均值在 } 66.8 \text{ 英寸和 } 68.3 \text{ 英寸之间的样本比例} \\
 &= (\text{正态曲线下 } z = -2.0 \text{ 和 } z = 0.5 \text{ 之间的面积}) \\
 &= (z = -2.0 \text{ 和 } z = 0 \text{ 之间的面积}) + (z = 0 \text{ 和 } z = 0.5 \text{ 之间的面积}) \\
 &= 0.4772 + 0.1915 = 0.6687
 \end{aligned}$$

则预计的样本数为 $80 \times 0.6687 = 53.496$, 或 53.

(b)

$$66.4 \text{ 的标准值} = \frac{66.4 - 68.0}{0.6} = -2.67$$

如图 8-1(b) 所示

$$\begin{aligned}
 & \text{均值小于 } 66.4 \text{ 英寸的样本比例} \\
 &= (\text{正态曲线下 } z = -2.67 \text{ 左方的面积}) \\
 &= (z = 0 \text{ 左方的面积}) - (z = -2.67 \text{ 和 } z = 0 \text{ 之间的面积}) \\
 &= 0.5 - 0.4962 = 0.0038
 \end{aligned}$$

则预计的样本数为 $80 \times 0.0038 = 0.304$, 或 0.

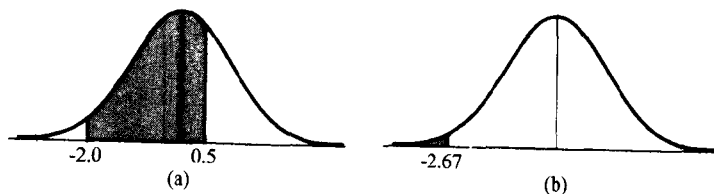


图 8-1

- 8.5 500 个轴承滚珠的均值的均值为 5.02 克, 标准差为 0.30 克. 从中随机抽取 100 个轴承滚珠, 求下列事件发生的概率: (a) 总重量在 496 克和 500 克之间, (b) 总重量大于 510 克.

解 对于均值的抽样分布, $\mu_{\bar{X}} = \mu = 5.02$ 克, 且

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} = \frac{0.30}{\sqrt{100}} \sqrt{\frac{500 - 100}{500 - 1}} = 0.027 \text{ 克}$$

(a) 如果 100 个轴承滚珠的平均重量在 4.96 克和 5.00 克之间, 那么它们的总重量在 496 克和 500 克之间.

$$4.96 \text{ 的标准值} = \frac{4.96 - 5.02}{0.027} = -2.22$$

$$5.00 \text{ 的标准值} = \frac{5.00 - 5.02}{0.027} = -0.74$$

如图 8-2(a) 所示

$$\begin{aligned}
 & \text{要求的概率} = (z = -2.22 \text{ 和 } z = -0.74 \text{ 之间的面积}) \\
 &= (z = -2.22 \text{ 和 } z = 0 \text{ 之间的面积}) \\
 &\quad - (z = -0.74 \text{ 和 } z = 0 \text{ 之间的面积}) \\
 &= 0.4868 - 0.2704 = 0.2164
 \end{aligned}$$

(b) 如果 100 个轴承滚珠的平均重量超过 5.10 克, 那么它们的总重量将超过 510 克.

$$5.10 \text{ 的标准值} = \frac{5.10 - 5.02}{0.027} = 2.96$$

如图 8-2(b) 所示

$$\begin{aligned}
 & \text{要求的概率} = (z = 2.96 \text{ 右方的面积}) \\
 &= (z = 0 \text{ 右方的面积}) - (z = 0 \text{ 和 } z = 2.96 \text{ 之间的面积}) \\
 &= 0.5 - 0.4985 = 0.0015
 \end{aligned}$$

因此, 抽取含有 100 个轴承滚珠的样本 2000 个, 只有 3 个样本的总重量超过 510 克.

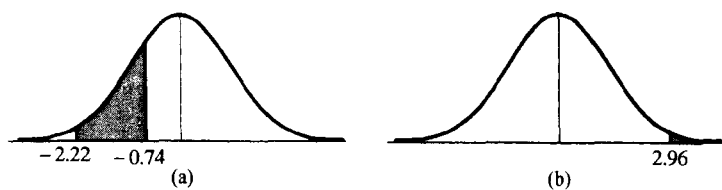


图 8-2

- 8.6 (a) 如何用随机数表从表 2.1 中选取 30 个随机样本? 假设每个样本包括 4 个学生且抽取是有放回的.
- (b) 求(a)中样本均值抽样分布的均值和标准差.
- (c) 将(b)中的结果和理论值相比较,并解释和理论不符合的地方.

解 (a) 用两位数给 100 个学生标号:00, 01, 02, ..., 99(见表 8.2). 则 5 个身高为 60~62 英寸的学生分别记为 00~04, 18 个身高为 63~65 英寸的学生分别记为 05~22 等. 每个学生的号码叫做**抽样号码**.

表 8.2

身高(英寸)	频数	抽样号码
60~62	5	00~04
63~65	18	05~22
66~68	42	23~64
69~71	27	65~91
72~74	8	92~99

我们从随机数表中抽取数字(附录 IX). 在第一行我们找到数列 51, 77, 27, 46, 40 等, 把它们作为随机抽样号码, 每个号码对应一个它所表示的学生的身高. 如 51 对应于一个 66 英寸~68 英寸的身高, 可视为 67 英寸. 类似地, 77, 27 和 46 分别对应于 70 英寸, 67 英寸和 67 英寸的身高.

这样, 我们得到表 8.3, 表中列出了抽取的抽样号码、对应的身高和每个样本的平均身高. 要指出的是, 虽然我们在第一行就已经引用了随机数表, 我们仍可以采用别的方式另取其他数.

表 8.3

抽取的抽样号码	相应的身高	平均身高	抽取的抽样号码	相应的身高	平均身高
1. 51, 77, 27, 46	67, 70, 67, 67	67.75	16. 11, 64, 55, 58	64, 67, 67, 67	66.25
2. 40, 42, 33, 12	67, 67, 67, 64	66.25	17. 70, 56, 97, 43	70, 67, 73, 67	69.25
3. 90, 44, 46, 62	70, 67, 67, 67	67.75	18. 74, 28, 93, 50	70, 67, 73, 67	69.25
4. 16, 28, 98, 93	64, 67, 73, 73	69.25	19. 79, 42, 71, 30	70, 67, 70, 67	68.50
5. 58, 20, 41, 86	67, 64, 67, 70	67.00	20. 58, 60, 21, 33	67, 67, 64, 67	66.25
6. 19, 64, 08, 70	64, 67, 64, 70	66.25	21. 75, 79, 74, 54	70, 70, 70, 67	69.25
7. 56, 24, 03, 32	67, 67, 61, 67	65.50	22. 06, 31, 04, 18	64, 67, 61, 64	64.00
8. 34, 91, 83, 58	67, 70, 70, 67	68.50	23. 67, 07, 12, 97	70, 64, 64, 73	67.75
9. 70, 65, 68, 21	70, 70, 70, 64	68.50	24. 31, 71, 69, 88	67, 70, 70, 70	69.25
10. 96, 02, 13, 87	73, 61, 64, 70	67.00	25. 11, 64, 21, 87	64, 67, 64, 70	66.25
11. 76, 10, 51, 08	70, 64, 67, 64	66.25	26. 03, 58, 57, 93	61, 67, 67, 73	67.00
12. 63, 97, 45, 39	67, 73, 67, 67	68.50	27. 53, 81, 93, 88	67, 70, 73, 70	70.00
13. 05, 81, 45, 93	64, 70, 67, 73	68.50	28. 23, 22, 96, 79	67, 64, 73, 70	68.50
14. 96, 01, 73, 52	73, 61, 70, 67	67.75	29. 98, 56, 59, 36	73, 67, 67, 67	68.50
15. 07, 82, 54, 24	64, 70, 67, 67	67.00	30. 08, 15, 08, 84	64, 64, 64, 70	65.50

(b) 表 8.4 给出了(a)中得到的身高的样本均值的频数分布,这就是**样本均值的抽样分布**.求均值和标准差的方法仍是第三章和第四章惯用的方法:

$$\begin{aligned}\text{均值} &= A + c\bar{u} = A + \frac{c \sum fu}{N} = 67.00 + \frac{0.75 \times 23}{30} = 67.58 \text{ 英寸} \\ \text{标准差} &= c \sqrt{u^2 - \bar{u}^2} = c \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} \\ &= 0.75 \sqrt{\frac{123}{30} - \left(\frac{23}{30}\right)^2} = 1.41 \text{ 英寸}\end{aligned}$$

表 8.4

样本均值	记数	f	u	fu	fu^2
64.00	—	1	-4	-4	16
64.75		0	-3	0	0
65.50	┐	2	-2	-4	8
66.25	正—	6	-1	-6	6
A → 67.00	正	4	0	0	0
67.75	正	4	1	4	4
68.50	正┐	7	2	14	28
69.25	正	5	3	15	45
70.00	—	1	4	4	16
		$\sum f = N = 30$		$\sum fu = 23$	$\sum fu^2 = 123$

(c) 样本均值的抽样分布的均值 $\mu_{\bar{X}}$ 应该等于总体的均值 μ , 即 67.45 英寸(见习题 3.22), 这和(b)中求出的值 67.58 英寸相一致.

样本均值的抽样分布的标准差(标准误差) $\sigma_{\bar{X}}$ 应该等于 σ/\sqrt{N} , 其中总体的标准差 $\sigma = 2.92$ 英寸(见习题 4.17), 样本的容量 $N = 4$. 则 $\sigma/\sqrt{N} = 2.92/\sqrt{4} = 1.46$ 英寸, 和(b)中的值 1.41 英寸相一致. 结果略有出入是因为只抽取了 30 个样本且样本容量(本题中为 4)太小.

比例的抽样分布

8.7 抛掷一枚均匀硬币 120 次, 求下列事件发生的概率: (a) 出现正面的次数占 40% 到 60%, (b) 出现正面的次数占 $\frac{5}{8}$ 或更多.

解 **解法一** 我们把抛掷硬币 120 次看作是从抛掷一枚硬币所有可能的结果这一无限总体中抽取的一个样本. 在总体中出现正面的概率为 $p = \frac{1}{2}$, 出现反面的概率为 $q = 1 - p = \frac{1}{2}$.

(a) 我们要求的是在 120 次中正面出现的次数在(120 的 40%)48 和(120 的 60%)72 之间的概率. 同第七章, 我们用二项分布的正态逼近来求此概率. 因为正面出现的次数是一离散型变量, 我们改求正面出现的次数在 47.5 和 72.5 之间的概率.

$$\mu = \text{正面出现次数的期望值} = Np = 120 \times \frac{1}{2} = 60$$

$$\sigma = \sqrt{Npq} = \sqrt{120 \times \frac{1}{2} \times \frac{1}{2}} = 5.48$$

$$47.5 \text{ 的标准值} = \frac{47.5 - 60}{5.48} = -2.28$$

$$72.5 \text{ 的标准值} = \frac{72.5 - 60}{5.48} = 2.28$$

如图 8-3,

$$\begin{aligned} \text{要求的概率} &= (\text{正态曲线下 } z = -2.28 \text{ 和 } z = 2.28 \text{ 之间的面积}) \\ &= 2(z = 0 \text{ 和 } z = 2.28 \text{ 之间的面积}) \\ &= 2 \times 0.4887 = 0.9774 \end{aligned}$$

解法二

$$\mu_P = p = \frac{1}{2} = 0.50$$

$$\sigma_P = \sqrt{\frac{pq}{N}} = \sqrt{\frac{\frac{1}{2} \times \frac{1}{2}}{120}} = 0.0456$$

$$40\% \text{ 的标准值} = \frac{0.40 - 0.50}{0.0456} = -2.19$$

$$60\% \text{ 的标准值} = \frac{0.60 - 0.50}{0.0456} = 2.19$$

$$\begin{aligned} \text{要求的概率} &= (\text{正态曲线下 } z = -2.19 \text{ 和 } z = 2.19 \text{ 之间的面积}) \\ &= 2 \times 0.4857 = 0.9714 \end{aligned}$$

即使这一结果已精确到两位有效数字,但仍不能完全准确,因为我们并没有用到比例实际上是离散型变量这一事实.为了弥补,我们用 0.40 减 $\frac{1}{2N}$, 0.60 加 $\frac{1}{2N}$; 因为 $N = 120$, $\frac{1}{2N} = 1/240 = 0.00417$, 则要求的比例的标准值分别为

$$\frac{0.40 - 0.00417 - 0.50}{0.0456} = -2.28 \text{ 和 } \frac{0.60 + 0.00417 - 0.50}{0.0456} = 2.28$$

这样就和解法一中的结果一样了.

注意, 0.40 - 0.00417 和 0.60 + 0.00417 对应于解法一中的比例 47.5/120 和 72.5/120.

(b) 用(a)中解法二的方法, 由 $\frac{5}{8} = 0.6250$

$$(0.6250 - 0.00417) \text{ 的标准值} = \frac{0.6250 - 0.00417 - 0.50}{0.0456} = 2.65$$

$$\begin{aligned} \text{要求的概率} &= (\text{正态曲线下 } z = 2.65 \text{ 右方的面积}) \\ &= (z = 0 \text{ 右方的面积}) - (z = 0 \text{ 和 } z = 2.65 \text{ 之间的面积}) \\ &= 0.5 - 0.4960 = 0.0040 \end{aligned}$$

8.8 有 500 个人, 每个人抛掷一枚均匀硬币 120 次. 预计有多少人能得到: (a) 出现正面的次数占 40% 到 60%, (b) 正面的次数占 $\frac{5}{8}$ 或更多?

解 这个问题和习题 8.7 密切相关. 我们可看成从抛掷一枚硬币的所有结果这一无限总体中抽取 500 个容量为 120 的样本.

(a) 习题 8.7 中的 (a) 指出: 如果每个样本包括抛掷一枚均匀硬币 120 次, 那么在所有可能的样本中, 出现正面的次数在 40% 和 60% 之间的占 97.74%. 因此, 在 500 个样本中, 我们预计大约能找到 (500 的 97.74%) 489 个有如此特征的样本. 所以会有 489 个人报告说, 他们的试验结果出现正面的次数占 40% 到 60%.

有趣的是, 预计会有 $500 - 489 = 11$ 个人报告说, 出现正面的次数的百分比不在 40% 和 60% 之间. 虽然这些人的硬币是均匀的, 但他们也有理由认为他们的硬币是负重的. 这种类型的错误正是概率中永远存在的风险.

(b) 和 (a) 中一样, 我们可断定大约有 $500 \times 0.0040 = 2$ 个人报告出现正面的次数占 $\frac{5}{8}$ 或更多.

8.9 已知某台机器生产的工具中有 2% 是次品. 装运 400 个这样的工具, 求下列事件发生的概率: (a) 次品不少于 3%, (b) 次品不多于 2%.

解

$$\mu_P = p = 0.02 \text{ 且 } \sigma_P = \sqrt{\frac{pq}{N}} = \sqrt{\frac{0.02 \times 0.98}{400}} = \frac{0.14}{20} = 0.007$$

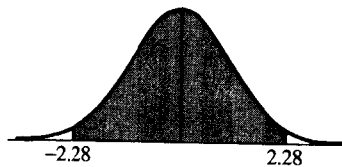


图 8-3

(a)解法一 对离散型随机变量进行修正, $1/(2N) = 1/800 = 0.00125$, 则有

$$(0.03 - 0.00125) \text{ 的标准值} = \frac{0.03 - 0.00125 - 0.02}{0.007} = 1.25$$

要求的概率 = (正态曲线下 $z = 1.25$ 右方的面积) = 0.1056

如果我们没有进行修正, 结果将是 0.0764.

解法二 有 400 的 3% = 12 个次品. 从连续的角度来考虑, 12 个或更多个工具就是 11.5 个或更多个.

$$\mu = Np = 400 \times 2\% = 8 \text{ 且 } \sigma = \sqrt{Npq} = \sqrt{400 \times 0.02 \times 0.98} = 2.8$$

则 11.5 的标准值 = $(11.5 - 8)/2.8 = 1.25$. 同上, 要求的概率是 0.1056.

(b)

$$(0.02 + 0.00125) \text{ 的标准值} = \frac{0.02 + 0.00125 - 0.02}{0.007} = 0.18$$

要求的概率 = (正态曲线下 $z = 0.18$ 左方的面积)

$$= 0.5000 + 0.0714 = 0.5714$$

如果我们没有进行修正, 结果将是 0.5000. 也可用(a)中解法二的方法.

8.10 某选举区的选举结果表明某一位候选人得到了 46% 的选票. 从选民中随机抽取 (a) 200, (b) 1000 人作民意测验, 求大多数人支持这位候选人的概率.

解 (a) $\mu_p = p = 0.46$ 且 $\sigma_p = \sqrt{\frac{pq}{N}} = \sqrt{\frac{0.46 \times 0.54}{200}} = 0.0352$

因为 $1/(2N) = 1/400 = 0.0025$, 所以样本中的大多数即指支持这位候选人的选民比例为 $0.50 + 0.0025 = 0.5025$ 或更多. (这个比例也可这样求得: 101 人或更多就是 200 人中的大多数, 如果将其视为连续型变量就是 100.5 人, 因此比例是 $100.5/200 = 0.5025$.)

$$0.5025 \text{ 的标准值} = \frac{0.5025 - 0.46}{0.0352} = 1.21$$

要求的概率 = (正态曲线下 $z = 1.21$ 右方的面积)

$$= 0.5000 - 0.3869 = 0.1131$$

$$(b) \mu_p = p = 0.46 \text{ 且 } \sigma_p = \sqrt{\frac{pq}{N}} = \sqrt{\frac{0.46 \times 0.54}{1000}} = 0.0158$$

$$0.5005 \text{ 的标准值} = \frac{0.5005 - 0.46}{0.0158} = 2.56$$

要求的概率 = (正态曲线下 $z = 2.56$ 右方的面积)

$$= 0.5000 - 0.4948 = 0.0052$$

差与和的抽样分布

8.11 变量 U_1 表示总体 3, 7, 8 中的任一元素, U_2 表示总体 2, 4 中的任一元素. 计算: (a) μ_{U_1} , (b) μ_{U_2} , (c) $\mu_{U_1 - U_2}$, (d) σ_{U_1} , (e) σ_{U_2} , (f) $\sigma_{U_1 - U_2}$.

解 (a) μ_{U_1} = 总体 U_1 的均值 = $\frac{1}{3}(3 + 7 + 8) = 6$

(b) μ_{U_2} = 总体 U_2 的均值 = $\frac{1}{2}(2 + 4) = 3$

(c) 总体包括 U_1 中的任意数和 U_2 中的任意数之差:

$$\begin{array}{ccccc} 3-2 & 7-2 & 8-2 & & 1 & 5 & 6 \\ 3-4 & 7-4 & 8-4 & \text{或} & -1 & 3 & 4 \end{array}$$

则

$$\mu_{U_1 - U_2} = (U_1 - U_2) \text{ 的均值} = \frac{1+5+6+(-1)+3+4}{6} = 3$$

由(a)和(b)可见, 一般的结果为 $\mu_{U_1 - U_2} = \mu_{U_1} - \mu_{U_2}$.

(d) $\sigma_{U_1}^2$ = 总体 U_1 的方差 = $\frac{(3-6)^2 + (7-6)^2 + (8-6)^2}{3} = \frac{14}{3}$ 或 $\sigma_{U_1} = \sqrt{\frac{14}{3}}$.

(e) $\sigma_{U_2}^2$ = 总体 U_2 的方差 = $\frac{(2-3)^2 + (4-3)^2}{2} = 1$ 或 $\sigma_{U_2} = 1$.

(f)

$$\sigma_{U_1 - U_2}^2 = \text{总体}(U_1 - U_2) \text{ 的方差}$$

$$= \frac{(1-3)^2 + (5-3)^2 + (6-3)^2 + (-1-3)^2 + (3-3)^2 + (4-3)^2}{6}$$

$$= \frac{17}{3}$$

或 $\sigma_{U_1 - U_2} = \sqrt{\frac{17}{3}}$. 由(d)和(e)可见, 一般的结果为 $\sigma_{U_1 - U_2} = \sqrt{\sigma_{U_1}^2 + \sigma_{U_2}^2}$.

- 8.12** A 品牌电灯泡的平均寿命为 1400 小时, 标准差为 200 小时. B 品牌电灯泡的平均寿命为 1200 小时, 标准差为 100 小时. 从每个品牌中随机抽取 125 个灯泡检验, 求 A 品牌灯泡的平均寿命比 B 品牌灯泡的平均寿命至少多 (a) 160 小时, (b) 250 小时的概率.

解 设 \bar{X}_A 和 \bar{X}_B 分别表示样本 A 和 B 的平均寿命. 则

$$\mu_{\bar{X}_A - \bar{X}_B} = \mu_{\bar{X}_A} - \mu_{\bar{X}_B} = 1400 - 1200 = 200 \text{ 小时}$$

$$\sigma_{\bar{X}_A - \bar{X}_B} = \sqrt{\frac{\sigma_A^2}{N_A} + \frac{\sigma_B^2}{N_B}} = \sqrt{\frac{(100)^2}{125} + \frac{(200)^2}{125}} = 20 \text{ 小时}$$

均值之差的标准化变量为

$$z = \frac{(\bar{X}_A - \bar{X}_B) - \mu_{\bar{X}_A - \bar{X}_B}}{\sigma_{\bar{X}_A - \bar{X}_B}} = \frac{(\bar{X}_A - \bar{X}_B) - 200}{20}$$

很接近于正态分布.

(a) 差 160 的标准值是 $(160 - 200)/20 = -2$. 则

$$\begin{aligned} \text{要求的概率} &= (\text{正态曲线下 } z = -2 \text{ 右方的面积}) \\ &= 0.5000 + 0.4772 = 0.9772 \end{aligned}$$

(b) 差 250 的标准值是 $(250 - 200)/20 = 2.50$. 则

$$\begin{aligned} \text{要求的概率} &= (\text{正态曲线下 } z = 2.50 \text{ 右方的面积}) \\ &= 0.5000 - 0.4938 = 0.0062 \end{aligned}$$

- 8.13** 某一品牌的轴承滚珠重 0.50 克, 标准差为 0.02 克. 现有两批该种轴承滚珠各 1000 个, 求两批轴承滚珠重量之差多于 2 克的概率.

解 设 \bar{X}_1 和 \bar{X}_2 分别表示两批轴承重量的均值. 则

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = 0.50 - 0.50 = 0$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{(0.02)^2}{1000} + \frac{(0.02)^2}{1000}} = 0.000895$$

均值之差的标准化变量为

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{0.000895}$$

很接近于正态分布.

两批轴承之间重量差为 2 克相当于均值差为 $2/1000 = 0.002$ 克. 或者是 $\bar{X}_1 - \bar{X}_2 \geq 0.002$, 或者是 $\bar{X}_1 - \bar{X}_2 \leq -0.002$, 即

$$z \geq \frac{0.002 - 0}{0.000895} = 2.23 \text{ 或 } z \leq \frac{-0.002 - 0}{0.000895} = -2.23$$

则 $P(z \geq 2.23 \text{ 或 } z \leq -2.23) = P(z \geq 2.23) + P(z \leq -2.23) = 2 \times (0.5000 - 0.4871) = 0.0258$.

- 8.14** A 和 B 玩“正面和反面”的游戏. 每人抛掷一枚硬币 50 次, 如果 A 得到的正面次数比 B 得到的正面次数多 5 或更多, 则 A 获胜, 否则, B 获胜. 求在任一场游戏中 A 不获胜的成败比.

解 用 P_A 和 P_B 分别表示 A 和 B 得到正面次数的比例. 如果我们假定硬币是均匀的, 则出现正面的概率 p 为 $\frac{1}{2}$. 故有

$$\mu_{P_A - P_B} = \mu_{P_A} - \mu_{P_B} = 0$$

$$\sigma_{P_A - P_B} = \sqrt{\sigma_{P_A}^2 + \sigma_{P_B}^2} = \sqrt{\frac{pq}{N_A} + \frac{pq}{N_B}} = \sqrt{2 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{50}} = 0.10$$

比例之差的标准化变量为 $z = (P_A - P_B - 0)/0.10$.

从连续的角度来考虑, 5 次或更多次出现正面即指 4.5 次或更多次出现正面, 所以比例之差应为 $4.5/50 = 0.09$ 或更多; 也就是说, 子大于或等于 $(0.09 - 0)/0.10 = 0.9$ (或 $z \geq 0.9$). 这一概率等于正态曲线下 $z = 0.9$ 右方的面积, 即为 $(0.5000 - 0.3159) = 0.1841$.

则 A 不获胜的成败比为 $(1 - 0.1841):0.1841 = 0.8159:0.1841$, 或 4.43:1.

- 8.15 两段距离测量为 27.3 厘米和 15.6 厘米, 标准差(标准误差)分别是 0.16 厘米和 0.08 厘米. 求两段距离的(a) 和(b) 差的均值和标准差.

解 如果两段距离分别记为 D_1 和 D_2 , 则

$$(a) \quad \mu_{D_1 + D_2} = \mu_{D_1} + \mu_{D_2} = 27.3 + 15.6 = 42.9 \text{ 厘米}$$

$$\sigma_{D_1 + D_2} = \sqrt{\sigma_{D_1}^2 + \sigma_{D_2}^2} = \sqrt{(0.16)^2 + (0.08)^2} = 0.18 \text{ 厘米}$$

$$(b) \quad \mu_{D_1 - D_2} = \mu_{D_1} - \mu_{D_2} = 27.3 - 15.6 = 11.7 \text{ 厘米}$$

$$\sigma_{D_1 - D_2} = \sqrt{\sigma_{D_1}^2 + \sigma_{D_2}^2} = \sqrt{(0.16)^2 + (0.08)^2} = 0.18 \text{ 厘米}$$

- 8.16 某种灯泡的平均寿命为 1500 小时, 标准差为 150 小时. 有三个灯泡串联在一起, 当一个灯泡损坏后, 另一个接着使用. 假设灯泡的寿命服从正态分布, 求三个灯泡总发光时间满足下列条件的概率: (a) 至少 5000 小时, (b) 至多 4200 小时.

解 假设灯泡的寿命分别为 L_1, L_2 和 L_3 . 则

$$\mu_{L_1 + L_2 + L_3} = \mu_{L_1} + \mu_{L_2} + \mu_{L_3} = 1500 + 1500 + 1500 = 4500 \text{ 小时}$$

$$\sigma_{L_1 + L_2 + L_3} = \sqrt{\sigma_{L_1}^2 + \sigma_{L_2}^2 + \sigma_{L_3}^2} = \sqrt{3 \times 150^2} = 260 \text{ 小时}$$

$$(a) \quad 5000 \text{ 的标准值} = \frac{5000 - 4500}{260} = 1.92$$

$$\begin{aligned} \text{要求的概率} &= (\text{正态曲线下 } z = 1.92 \text{ 右方的面积}) \\ &= 0.5000 - 0.4726 = 0.0274 \end{aligned}$$

$$(b) \quad 4200 \text{ 的标准值} = \frac{4200 - 4500}{260} = -1.15$$

$$\begin{aligned} \text{要求的概率} &= (\text{正态曲线下 } z = -1.15 \text{ 左方的面积}) \\ &= 0.5000 - 0.3749 = 0.1251 \end{aligned}$$

杂题

- 8.17 参考习题 8.1, 求样本(a) 方差抽样分布的均值, (b) 样本方差抽样分布的标准差(即样本方差的标准误差).

解 (a) 对应于习题 8.1 中的 25 个样本的样本方差是

0	0.25	4.00	9.00	20.25
0.25	0	2.25	6.25	16.00
4.00	2.25	0	1.00	6.25
9.00	6.25	1.00	0	2.25
20.25	16.00	6.25	2.25	0

样本方差抽样分布的均值是

$$\mu_{s^2} = \frac{25 \text{ 个样本的样本方差之和}}{25} = \frac{135}{25} = 5.40$$

因为 $N = 2$ 且 $\sigma^2 = 10.8$ (见习题 8.1(b)), 而 $(N - 1)\sigma^2/N = \frac{1}{2} \times 10.8 = 5.4$, 所以 $\mu_{s^2} = (N - 1)\sigma^2/N$.

结果表明有必要定义一个修正的样本方差

$$\hat{s}^2 = \frac{N}{N-1} s^2$$

则有 $\mu_s = \sigma^2$ (见 69 页的注释). 要注意的是总体方差的定义同前, 只有样本方差需要修正.

(b) 样本方差抽样分布的方差等于用上表中的每个数减去均值 5.40, 再将所得 25 个数的平方和除以 25. 因此 $\sigma_s^2 = 575.75/25 = 23.03$, 或 $\sigma_s = 4.80$.

8.18 如果抽样是无放回的, 求解习题 8.17.

解 (a) 在习题 8.17(a) 的表中, 位于零对角线的上方(或下方)共有 10 个数字给出了 10 个样本的方差. 则

$$\begin{aligned}\mu_s^2 &= \frac{0.25 + 4.00 + 9.00 + 20.25 + 2.25 + 6.25 + 16.00 + 1.00 + 6.25 + 2.25}{10} \\ &= 6.75\end{aligned}$$

这是一般结果

$$\mu_s^2 = \left(\frac{N_p}{N_p - 1} \right) \left(\frac{N-1}{N} \right) \sigma^2$$

的一个特殊情形, 将 $N_p = 5$, $N = 2$ 和 $\sigma^2 = 10.8$ 代入上式右边, 得 $\mu_s^2 = \frac{5}{4} \times \frac{1}{2} \times 10.8 = 6.75$.

(b) 将习题 8.17(a) 的表中的位于零对角线上方的 10 个数字各减去 6.75, 所得结果的平方和除以 10, 可得 $\sigma_s^2 = 39.675$, 或 $\sigma_s = 6.30$.

8.19 一大群学生体重的标准差为 10.0 磅. 从总体中抽取包含有 200 名学生的样本, 计算体重的样本标准差. 求样本标准差抽样分布的 (a) 均值, (b) 标准差.

解 我们可把样本看作是从无限总体中抽取或从有限总体中有放回地抽取得到的. 由表 8.1, 我们有

(a) 样本标准差抽样分布的均值为 $\mu_s = \sigma = 10.0$ 磅.

(b) 样本标准差抽样分布的标准差为 $\sigma_s = \sigma / \sqrt{2N} = 10 / \sqrt{400} = 0.50$ 磅.

8.20 在习题 8.19 中, 求标准差满足下列条件的样本的百分比: (a) 大于 11.0 磅, (b) 小于 8.8 磅.

解 样本标准差抽样分布近似于均值为 10.0 磅, 方差为 0.50 磅的正态分布.

(a) 11.0 的标准值是 $(11.0 - 10.0) / 0.50 = 2.0$. 正态曲线下 $z = 2.0$ 右方的面积是 $0.5 - 0.4772 = 0.0228$, 则要求的百分比是 2.3%.

(b) 8.8 的标准值是 $(8.8 - 10.0) / 0.50 = -2.4$. 正态曲线下 $z = -2.4$ 左方的面积是 $0.5 - 0.4918 = 0.0082$, 则要求的百分比是 0.8%.

补充习题

均值的抽样分布

8.21 一个总体包含 4 个数字 3, 7, 11 和 15. 考察从中有放回地抽取出的所有容量为 2 的样本. 求 (a) 总体均值, (b) 总体标准差, (c) 样本均值抽样分布的均值, (d) 样本均值抽样分布的标准差. 运用合适的公式由 (a) 和 (b) 验证 (c) 和 (d).

8.22 如果抽取是无放回的, 求解习题 8.21.

8.23 1500 个轴承滚珠的质量服从均值为 22.40 克, 标准差为 0.048 克的正态分布. 从中随机抽取 300 个容量为 36 的样本, 求样本均值抽样分布的期望的均值和标准差, 如果 (a) 抽取是有放回的, (b) 抽取是无放回的.

8.24 如果总体包含有 72 个轴承滚珠, 求解习题 8.23.

8.25 在习题 8.23 中, 有多少个随机样本的均值满足下列条件: (a) 在 22.39 克和 22.41 克之间, (b) 大于 22.42 克, (c) 小于 22.37 克, (d) 小于 22.38 克或大于 22.41 克.

8.26 某公司生产的电子管的平均寿命为 800 小时, 标准差为 60 小时. 从中随机抽取 16 支电子管, 求其平均寿命满足下列条件的概率: (a) 在 790 小时和 810 小时之间, (b) 小于 785 小时, (c) 大于 820 小时, (d) 在 770 小时和 830 小时之间.

- 8.27 如果从总体中抽取了 64 支电子管, 求解习题 8.26, 并解释两题的区别.
- 8.28 某百货商店收到的包裹重量的均值为 300 磅, 标准差为 50 磅, 如果电梯的安全极限为 8200 磅, 求随机收到的 25 个包裹放入电梯超重的概率.

随机数表

- 8.29 有放回地选取 (a) 15, (b) 30, (c) 45, (d) 60 个容量为 4 的样本, 用另一组随机数求解习题 8.6, 并比较每一种情形中的理论上的结果.
- 8.30 如果有放回抽取的样本容量不是 4, 而是 (a) 2, (b) 8, 求解习题 8.29.
- 8.31 如果抽样是无放回地, 求解习题 8.6, 并和理论上的结果相比较.
- 8.32 (a) 说明如何从习题 3.61 的分布中选取 30 个容量为 2 的样本.
(b) 求样本均值的抽样分布的均值和标准差, 并和理论上的结果相比较.
- 8.33 如果样本容量为 4, 求解习题 8.32.

比例的抽样分布

- 8.34 假定男孩和女孩出生的概率相等, 在 200 个将要出生的孩子中, 求下列事件发生的概率: (a) 男孩少于 40%, (b) 女孩在 43% 和 57% 之间, (c) 男孩多于 54%.
- 8.35 在 1000 个各有 200 个孩子的样本中, 预计能找到多少个样本满足下列条件? (a) 男孩少于 40%, (b) 女孩在 40% 和 60% 之间, (c) 女孩不少于 53%.
- 8.36 如果每个样本含有 100 个孩子, 求解习题 8.34, 并解释两题结果的差别.
- 8.37 一个罐子中有 80 个球, 其中 60% 是红色的, 40% 是白色的. 从罐子中有放回地抽取 50 个含有 20 个球的样本, 预计有多少个样本含有 (a) 等量的红球和白球, (b) 12 个红球和 8 个白球, (c) 8 个红球和 12 个白球, (d) 10 个或更多的白球?
- 8.38 设计一个试验验证习题 8.37 的结果. 你可以按正确的比例用写有 R 和 W 的纸条分别代替红球和白球. 若用两种不同的硬币代替球会引起什么样的误差.
- 8.39 某工厂生产 1000 批灯泡, 每批 100 个. 如果灯泡中有 5% 是次品, 则预计能有多少批灯泡满足 (a) 正品少于 90 个, (b) 正品不少于 98 个?

差与和的抽样分布

- 8.40 A 和 B 两种品牌电缆的拉断力均值分别为 4000 磅和 4500 磅, 标准差分别为 300 磅和 200 磅. 检验 A 品牌的 100 条电缆和 B 品牌的 50 条电缆, 求 B 品牌电缆的拉断力均值满足下列条件的概率: (a) 至少比 A 的大 600 磅, (b) 至少比 A 的大 450 磅.
- 8.41 如果两种品牌的电缆都检验 100 条, 求解习题 8.40, 并解释两题的区别.
- 8.42 在一次智力测验中, 学生的均分是 72 分, 标准差是 8 分. 有两组学生, 各有 28 人和 36 人, 求它们均分之差满足下列条件的概率: (a) 不小于 3 分, (b) 不小于 6 分, (c) 在 2 分和 5 分之间.
- 8.43 一个罐子中有 60 个红球和 40 个白球. 从中有放回地抽取两组球, 每组 30 个, 并记下它们的颜色. 求两组球中的红球数之差不小于 8 的概率.
- 8.44 如果抽取每组球时, 抽取都是无放回的, 求解习题 8.43.
- 8.45 某选举区选举结果表明某位候选人获得了 65% 的选票. 随机抽取两组样本, 每组含有 200 名选民, 求两组样本中支持这位候选人的选民比例之差大于 10% 的概率.
- 8.46 如果 U_1 和 U_2 是习题 8.11 中的两组数字, 证明: (a) $\mu_{U_1+U_2} = \mu_{U_1} + \mu_{U_2}$,
(b) $\sigma_{U_1+U_2} = \sqrt{\sigma_{U_1}^2 + \sigma_{U_2}^2}$.
- 8.47 测得 3 个物体的重量分别为 20.48 克, 35.97 克和 62.34 克, 标准差分别为 0.21 克, 0.46 克和 0.54 克. 求重量之和的 (a) 均值, (b) 标准差.
- 8.48 一节电池的平均电压是 15.0 伏, 标准差是 0.2 伏. 求四节这样的电池串联后总电压不小于 60.8 伏的概率.

杂题

- 8.49 含有 7 个数字的总体的均值是 40, 标准差是 3. 从总体中随机抽取容量为 5 的样本, 求样本方差抽样分

布的均值,如果(a)抽取有放回,(b)抽取无放回.

- 8.50** 某公司生产的电子管的平均寿命是 900 小时,标准差是 80 小时.该公司生产了 1000 批电子管,每批有 100 个.预计能有多少批产品满足(a) 平均寿命超过 910 小时,(b) 寿命的标准差超过 95 小时? 必须要作出何种假设?
- 8.51** 如果习题 8.50 中寿命的中位数是 900 小时,预计能有多少批产品的寿命的中位数超过 910 小时? 和习题 8.50(a)中的答案相比较并解释结果.
- 8.52** 某次测验成绩服从均值为 72 分,方差为 8 分的正态分布.(a) 求得分在前 20% 的学生的最低分,(b) 随机抽取 100 名学生,求得分在前 20% 的学生的最低分小于 76 的概率.

第九章 统计估计理论

参数的估计

在上一章,我们了解到如何应用抽样理论来获取从已知总体中随机抽取得到的样本的信息.但是从实用的角度来看,如何通过抽取的样本来推断有关总体的信息才是更重要的.统计推断处理的就是这类问题,其中要用到抽样理论的一些准则.

统计推断中的一个重要问题是根据**样本统计量**或简称**统计量**(如样本均值和方差)得到相应的**总体参数**或简称**参数**(如总体均值和方差)的估计.本章就要讨论这一问题.

无偏估计

如果一个统计量的抽样分布的均值等于相应的总体参数,那么这个统计量就是此参数的一个**无偏估计量**;否则,就称为**有偏估计量**.统计量的相应值分别称为**无偏估计**或**有偏估计**.估计量也常简称为估计.

例 1 样本均值抽样分布的均值 $\mu_{\bar{X}}$ 等于总体的均值 μ , 则样本均值 \bar{X} 是总体均值 μ 的一个无偏估计.

例 2 样本方差抽样分布的均值为 $\mu_{s^2} = \frac{N-1}{N}\sigma^2$, 其中 σ^2 是总体方差, N 是样本的容量(见表 8.1), 则样本方差 s^2 是总体方差 σ^2 的一个有偏估计. 对于修正的样本方差

$$s^2 = \frac{N}{N-1}s^2$$

因为 $\mu_{s^2} = \sigma^2$, 所以 s^2 是 σ^2 的无偏估计. 但是, \bar{s} 是 σ 的有偏估计.

也可改用术语期望(见第六章)来叙述: 如果一个统计量的期望等于相应的总体参数, 则此统计量是无偏的. 由于 $E(\bar{X}) = \mu$, $E(s^2) = \sigma^2$, 故 \bar{X} 和 s^2 是无偏的.

有效估计

如果两个统计量的抽样分布有相同的均值(或期望), 那么方差较小的那个统计量称为此均值的**有效估计量**, 另一个称为**无效估计量**. 统计量的相应值称为**有效估计**或**无效估计**.

在均值的所有无偏估计量中, 方差最小的那个统计量常被称为此均值的**最有效估计量**或**最优估计量**.

例 3 样本均值和中位数的抽样分布有相同的均值, 都为总体均值. 但是, 样本均值抽样分布的方差要小于中位数抽样分布的方差(见表 8.1). 因此两者相比较, 样本均值是总体均值的有效估计, 而样本中位数是无效估计.

在总体均值的所有线性估计量中, 样本均值是最有效(或最优)估计量.

在实际问题中, 常常要用到无效估计, 因为它们有时计算起来相对较为简便.

点估计和区间估计

如果用一个数来估计总体的参数, 那么这种估计叫做参数的**点估计**. 如果给出两个数, 指出参数位于其间, 那么这种估计叫做参数的**区间估计**.

区间估计更加精确, 因而要优于点估计.

例 4 如果我们说测量的一段距离是 5.28 米, 我们就给出了一个点估计. 但如果我们说这段距离是 5.28 米 \pm 0.3 米(即距离在 5.25 米和 5.31 米之间), 我们就给出了一个区间估计.

总体参数的置信区间估计

用 μ_S 和 σ_S 分别表示统计量 S 的均值和标准差. 如果 S 的抽样分布是近似正态的(当样本容量 $N \geq 30$ 时对于许多统计量都成立), 则实际的一个样本统计量 S 位于区间 $(\mu_S - \sigma_S, \mu_S + \sigma_S)$, $(\mu_S - 2\sigma_S, \mu_S + 2\sigma_S)$ 和 $(\mu_S - 3\sigma_S, \mu_S + 3\sigma_S)$ 的可能性大约分别为 68.27%, 95.45% 和 99.73%.

同样地, 我们在区间 $(S - \sigma_S, S + \sigma_S)$, $(S - 2\sigma_S, S + 2\sigma_S)$ 和 $(S - 3\sigma_S, S + 3\sigma_S)$ 内找到 μ_S 的可能性大约分别为 68.27%, 95.45% 和 99.73%. 因此, 我们分别称这些区间为 μ_S 的 68.27%, 95.45% 和 99.73% 的**置信区间**. 这些区间的端点值 $(S \pm \sigma_S, S \pm 2\sigma_S$ 和 $S \pm 3\sigma_S)$ 则称为 68.27%, 95.45% 和 99.73% 的**置信界限**或**置信限**.

类似地, $S \pm 1.96\sigma_S$ 和 $S \pm 2.58\sigma_S$ 是 μ_S 的 95% 和 99% (0.95 或 0.99) 的置信界限. 其中, 百分数叫做**置信水平**, 数字 1.96, 2.58 等叫做**临界值**, 记为 z_c . 由置信水平我们能找到临界值, 反之亦然.

表 9.1 给出了对应于不同置信水平的常用的临界值 z_c . 对于表中未列出的置信水平, 相应的 z_c 值可由正态曲线面积表(见附录 II)求得.

表 9.1

置信水平	99.73%	99%	98%	96%	95.45%	95%	90%	80%	68.27%	50%
z_c	3.00	2.58	2.33	2.05	2.00	1.96	1.645	1.28	1.00	0.6745

均值的置信区间

如果统计量 S 指的是样本均值 \bar{X} , 则估计总体均值 μ 的 95% 和 99% 的置信限分别为 $\bar{X} \pm 1.96\sigma_{\bar{X}}$ 和 $\bar{X} \pm 2.58\sigma_{\bar{X}}$. 更一般地, 置信限为 $\bar{X} \pm z_c\sigma_{\bar{X}}$, 其中 z_c (依赖于不同的置信水平) 可由表 9.1 查得. 代入第八章中所求得的 $\sigma_{\bar{X}}$, 如果总体是无限的或总体有限但抽取是有放回的, 则总体均值的置信限为

$$\bar{X} \pm z_c \frac{\sigma}{\sqrt{N}} \quad (1)$$

如果抽取是无放回的且总体容量 N_p 有限, 则总体均值的置信限为

$$\bar{X} \pm z_c \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} \quad (2)$$

一般说来, 总体的标准差 σ 是未知的, 要得到上述置信限, 可用 σ 的估计 s 或 s 代替 σ . 当 $N \geq 30$ 时, 效果较好些. 当 $N < 30$ 时, 近似性较差, 就要用到小样本抽样理论(见第十一章).

比例的置信区间

如果统计量 S 指的是从一个服从二项分布且成功的比例(即成功的概率)是 p 的总体中抽取的容量为 N 的样本的成功比例 P , 则 p 的置信限是 $P \pm z_c\sigma_P$. 代入第八章中求得的 σ_P , 如果总体是无限的或总体有限但抽取是有放回的, 则总体比例的置信限为

$$P \pm z_c \sqrt{\frac{pq}{N}} = P \pm z_c \sqrt{\frac{p(1-p)}{N}} \quad (3)$$

如果抽取是无放回的且总体容量 N_p 有限, 则总体比例的置信限为

$$P \pm z_c \sqrt{\frac{pq}{N}} \sqrt{\frac{N_p - N}{N_p - 1}} \quad (4)$$

为了计算出置信限, 我们要用 P 来估计 p , 当 $N \geq 30$ 时, 效果较好些. 习题 9.12 中给出了一个求得置信限的更精确的方法.

差与和的置信区间

如果 S_1 和 S_2 是抽样分布近似正态的两个样本统计量且样本是独立的(见第八章),则对应于 S_1 和 S_2 的总体参数之差的置信限为

$$S_1 - S_2 \pm z_c \sigma_{S_1 - S_2} = S_1 - S_2 \pm z_c \sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2} \quad (5)$$

而总体参数之和的置信限为

$$S_1 + S_2 \pm z_c \sigma_{S_1 + S_2} = S_1 + S_2 \pm z_c \sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2} \quad (6)$$

例如,如果总体是无限的,两个总体均值之差的置信限为

$$\bar{X}_1 - \bar{X}_2 \pm z_c \sigma_{\bar{X}_1 - \bar{X}_2} = \bar{X}_1 - \bar{X}_2 \pm z_c \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \quad (7)$$

其中 \bar{X}_1, σ_1, N_1 和 \bar{X}_2, σ_2, N_2 分别表示从总体中抽取的两个样本的均值、总体标准差和样本容量.

类似地,如果总体是无限的,两个总体比例之差的置信限为

$$P_1 - P_2 \pm z_c \sigma_{P_1 - P_2} = P_1 - P_2 \pm z_c \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}} \quad (8)$$

其中 P_1 和 P_2 表示两个样本的比例, N_1 和 N_2 表示从总体中抽取的两个样本的容量, p_1 和 p_2 表示两个总体的比例(用 P_1 和 P_2 估计).

标准差的置信区间

服从正态分布的总体的标准差 σ 是用样本的标准差 s 估计的,由表 8.1, 它的置信限为

$$s \pm z_c \sigma_s = s \pm z_c \frac{\sigma}{\sqrt{2N}} \quad (9)$$

为求得置信限,用 s 或 s 估计 σ .

可能误差

对应于统计量 S 的总体参数的 50% 置信限为 $S \pm 0.6745\sigma_s$. 称 $0.6745\sigma_s$ 为估计的可能误差.

习题及解答

无偏估计与有效估计

9.1 给出满足下列条件的估计量的例子:(a) 无偏且有效,(b) 无偏但无效,(c) 有偏且无效.

解 (a) 样本均值 \bar{X} 和修正的样本方差

$$s^2 = \frac{N}{N-1} s^2$$

分别是总体均值和方差是满足条件的估计量.

(b) 样本中位数和样本统计量 $\frac{1}{2}(Q_1 + Q_3)$ 是满足条件的两个例子. 其中 Q_1 和 Q_3 分别是样本的第 1 个四分位数和第 3 个四分位数. 因为这两个统计量的抽样分布的均值都等于总体的均值, 所以它们都是总体均值的无偏估计.

(c) 样本标准差 s , 修正后的样本标准差 s , 平均偏差和半内四分距是满足条件的总体标准差的四个估计量.

9.2 一个样本包含了某位科学家测量到的一个球体直径的 5 次记录: 6.33, 6.37, 6.36, 6.32 和 6.37 厘米. 求(a) 真实均值, (b) 真实方差的无偏且有效的估计.

解 (a) 真实均值(即总体均值)的无偏且有效的估计为

$$\bar{X} = \frac{\sum X}{N} = \frac{6.33 + 6.37 + 6.36 + 6.32 + 6.37}{5} = 6.35 \text{ 厘米}$$

(b) 真实方差(即总体方差)的无偏且有效的估计为

$$\begin{aligned} s^2 &= \frac{N}{N-1} s^2 = \frac{\sum (X - \bar{X})^2}{N-1} \\ &= \frac{(6.33 - 6.35)^2 + (6.37 - 6.35)^2 + (6.36 - 6.35)^2 + (6.32 - 6.35)^2 + (6.37 - 6.35)^2}{5-1} \\ &= 0.00055 \text{ 厘米}^2 \end{aligned}$$

注意,即使 $s = \sqrt{0.00055} = 0.023$ 厘米是真实标准差的一个估计,但既不是无偏的也不是有效的.

9.3 假设 XYZ 大学里 100 名男生的身高是这所大学 1546 名男生身高的一个随机样本. 求

(a) 真实均值, (b) 真实方差的无偏且有效的估计.

解 (a) 由习题 3.22, 真实均值的无偏且有效的估计为 $\bar{X} = 67.45$ 英寸.

(b) 由习题 4.17, 真实方差的无偏且有效的估计为

$$s^2 = \frac{N}{N-1} s^2 = \frac{100}{99} \times 8.5275 = 8.6136$$

则 $s = \sqrt{8.6136} = 2.93$ 英寸. 注意, 因为 N 很大, 所以 s^2 和 s^2 或 s 和 s 在本质上没有什么区别.

注意, 我们并没有用到 Sheppard 的修正方法, 若要考虑到这一因素, 应有 $s = 2.79$ 英寸(见习题 4.21).

9.4 给出习题 9.2 中球体直径真实均值的一个无偏但无效的估计.

解 中位数是总体均值的一个无偏但无效的估计. 将五个测量值按大小排列, 中位数是 6.36 厘米.

均值的置信区间

9.5 求习题 9.3 中男生平均身高的(a) 95%, (b) 99%的置信区间.

解 (a) 95%的置信限是 $\bar{X} \pm 1.96\sigma/\sqrt{N}$. 由 $\bar{X} = 67.45$ 及 $s = 2.93$ 作为 σ 的估计(见习题 9.3), 置信限为 $67.45 \pm 1.96 \times (2.93/\sqrt{100})$ 或 67.45 ± 0.57 . 因此, 总体均值 μ 的 95%的置信区间是 66.88 英寸到 68.02 英寸, 记做 $66.88 < \mu < 68.02$.

因此我们可以说总体的平均身高在 66.88 英寸和 68.02 英寸之间的概率大约为 95% 或 0.95. 可记为 $P(66.88 < \mu < 68.02) = 0.95$. 这就相当于说我们有 95% 的把握断定总体均值或真实均值在 66.88 英寸和 68.02 英寸之间.

(b) 99%的置信限是 $\bar{X} \pm 2.58\sigma/\sqrt{N} = \bar{X} \pm 2.58s/\sqrt{N} = 67.45 \pm 2.58 \times (2.93/\sqrt{100}) = 67.45 \pm 0.76$. 因此, 总体均值 μ 的 99%的置信区间是 66.69 英寸到 68.21 英寸, 记做 $66.69 < \mu < 68.21$.

为了得到上述置信区间, 我们假定总体是无限的或总体很大, 这可以和有放回抽样等同考虑. 对于有限总体, 若抽样是无放回的, 则用

$$\frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} \quad \text{代替} \quad \frac{\sigma}{\sqrt{N}}$$

但是, 我们可将

$$\sqrt{\frac{N_p - N}{N_p - 1}} = \sqrt{\frac{1546 - 100}{1546 - 1}} = 0.967$$

看成 1.0, 因此并不需要作替换. 若要作替换, 上面的置信限分别变为 67.45 ± 0.56 英寸和 67.45 ± 0.73 英寸.

9.6 某农场有 5000 棵树有待砍伐. 从中随机选择 100 棵树并记下它们的高度. 以英寸为单位的高度列在表 9.2 中. 用 Minitab 建立 5000 棵树平均高度的 95% 置信区间. 如果每英尺树可卖 2.40 美元, 给出 5000 棵树价格的下界和上界.

解 下面用 Minitab 给出的置信区间表明 5000 棵树的平均高度最低可达 57.24 英寸, 最高可达 61.20 英寸. 则 5000 棵树的总长度在 $57.24 \times 5\,000 = 286\,200$ 英寸和 $61.20 \times 5\,000 = 306\,000$ 英寸之

间. 如果每英尺的树可卖 2.40 美元, 则每英寸可卖 0.2 美元. 所以 5000 棵树的价格以 95% 的置信水平在 $286000 \times 0.2 = 57\,200$ 美元和 $306000 \times 0.2 = 61200$ 美元之间.

表 9.2

56	61	52	62	63	34	47	35	44	59
70	61	65	51	65	72	55	71	57	75
53	48	55	67	60	60	73	74	43	74
71	53	78	59	56	62	48	65	68	51
73	62	80	53	64	44	67	45	58	48
50	57	72	55	56	62	72	57	49	62
46	61	52	46	72	56	46	48	57	52
54	73	71	70	66	67	58	71	75	50
44	59	56	54	63	43	68	69	55	63
48	49	70	60	67	47	49	69	66	73

Data Display

height

```

56 70 53 71 73 50 46 54 44
48 61 61 48 53 62 57 61 73
59 49 52 65 55 78 80 72 52
71 56 70 62 51 67 59 53 55
46 70 54 60 63 65 60 56 64
56 72 66 63 67 34 72 60 62
44 62 56 67 43 47 47 55 73
48 67 72 46 58 68 49 35 71
74 65 45 57 48 71 69 69 44
57 43 68 58 49 57 75 55 66
59 75 74 51 48 62 52 50 63
73

```

MTB> standard deviation cl

Column Standard Deviation

Standard Deviation of height = 10.111

MTB>zinterval 95 percent confidence sd = 10.111data in cl

Confidence Intervals

The assumed sigma = 10.1

Variable	N	Mean	StDev	SE	Mean	95.0% CI
Height	100	59.22	10.11	1.01	(57.24, 61.20)	

9.7 现对一些天主教神父作调查, 每位神父要上报在过去的一年里所主持的洗礼、婚礼和丧礼的总次数. 有关数据列在表 9.3 中. 用这些数据为过去一年里平均每位神父主持的洗礼、婚礼和丧礼的总次数 μ 建立 95% 的置信区间. 用置信区间公式和 Minitab 的 Zinterval 命令建立区间.

表 9.3

32	44	48	35	34	29	31	61	37	41
31	40	44	43	41	40	41	31	42	45
29	40	42	51	16	24	40	52	62	41
32	41	45	24	41	30	42	47	30	46
38	42	26	34	45	58	57	35	62	46

解 先将表 9.3 中的数据输入到 Minitab 工作页面的第 1 行, 并为其命名为 'number'. 然后输入 standard deviation 命令.

MTB>mean cl

Column Mean

Mean of Number = 40.261

MTB>standard deviation cl

Column Standard Deviation

Standard deviation of Number = 9.9895

均值的标准误差等于 $9.9895 / \sqrt{50} = 1.413$, 临界值是 1.96, 误差的 95% 边界是 $1.96 \times 1.413 = 2.769$. 所以置信区间为 $40.261 - 2.769 = 37.492$ 到 $40.261 + 2.769 = 43.030$.

输入 Zinterval 命令得到如下的结果:

MTB>Zinterval 95% confidence sd = 9.9895 data in cl

Zconfidence Intervals

The assumed sigma = 9.99

Variable N Mean StDev SE Mean 95.00% CI

Number 50 40.26 9.99 1.41 (37.49, 43.03)

我们可以有 95% 的把握确定所有神父主持次数的真实均值在 37.49 和 43.03 之间.

- 9.8** 生理学家在测量反应时间时, 估计的标准差是 0.05 秒. 为了有 (a) 95%, (b) 99% 的把握保证估计误差不超过 0.01 秒, 他必须要抽取多大的样本?

解 (a) 95% 的置信限是 $\bar{X} \pm 1.96\sigma/\sqrt{N}$, 估计误差是 $1.96\sigma/\sqrt{N}$. 令 $\sigma = 0.05$, 如果 $1.96 \times 0.05/\sqrt{N} = 0.01$, 那么误差等于 0.01, 即 $\sqrt{N} = 1.96 \times 0.05/0.01 = 9.8$ 或 $N = 96.04$. 因此, 我们可以有 95% 的把握确定: 如果 N 为 97 或更大, 则估计的误差小于 0.01 秒.

另解 若有

$$\frac{\sqrt{N}}{1.96 \times 0.05} \geq \frac{1}{0.01} \quad \text{或} \quad \sqrt{N} \geq \frac{1.96 \times 0.05}{0.01} = 9.8$$

则

$$\frac{1.96 \times 0.05}{\sqrt{N}} \leq 0.01.$$

所以 $N \geq 96.04$ 或 $N \geq 97$.

(b) 99% 的置信限是 $\bar{X} \pm 2.58\sigma/\sqrt{N}$, 则 $2.58 \times 0.05/\sqrt{N} = 0.01$ 或 $N = 166.4$. 因此, 我们可以有 99% 的把握确定: 只有当 N 为 167 或更大时, 估计的误差才会小于 0.01 秒.

- 9.9** 从 200 名学生中随机抽取的 50 名学生的数学平均分为 75 分, 标准差为 10 分.

(a) 200 名学生数学平均分的 95% 的置信限是多少?

(b) 我们可以在多大置信水平上说 200 名学生的平均分是 75 ± 1 ?

解 (a) 因为总体容量相对于样本容量来说并不是很大, 所以我们必须要作些调整. 则 95% 的置信限为

$$\bar{X} \pm 1.96\sigma_{\bar{X}} = \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} = 75 \pm 1.96 \frac{10}{\sqrt{50}} \sqrt{\frac{200 - 50}{200 - 1}} = 75 \pm 2.4$$

(b) 置信限可表示为

$$\bar{X} \pm z_c \sigma_{\bar{X}} = \bar{X} \pm z_c \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} = 75 \pm z_c \frac{10}{\sqrt{50}} \sqrt{\frac{200 - 50}{200 - 1}} = 75 \pm 1.23 z_c$$

而它必须等于 75 ± 1 , 则 $1.23 z_c = 1$ 或 $z_c = 0.81$. 正态曲线下 $z = 0$ 和 $z = 0.81$ 之间的面积为 0.2910, 因此要求的置信水平为 $2 \times 0.2910 = 0.582$ 或 58.2%.

比例的置信区间

- 9.10 对某一选举区内随机抽取的 100 位选民的民意调查表明他们中的 55% 支持某位候选人. 求所有选民中支持这位候选人的比例的 (a) 95%, (b) 99%, (c) 99.73% 的置信限.

解 总体比例 p 的 95% 的置信限为 $P \pm 1.96 \sigma_P = P \pm 1.96 \sqrt{P(1-P)/N} = 0.55 \pm 1.96 \sqrt{0.55 \times 0.45/100} = 0.55 \pm 0.10$, 其中我们用样本比例 P 来估计 p .

(b) p 的 99% 的置信限为 $0.55 \pm 2.58 \sqrt{0.55 \times 0.45/100} = 0.55 \pm 0.13$.

(c) p 的 99.73% 的置信限为 $0.55 \pm 3 \sqrt{0.55 \times 0.45/100} = 0.55 \pm 0.15$.

习题 9.12 提供了一个更为精确的方法.

- 9.11 在习题 9.10 中, 为了能有 (a) 95%, (b) 99.73% 的把握确定那位候选人必定当选, 要从选民中抽取多大的样本?

解 p 的置信限为 $P \pm z_c \sqrt{p(1-p)/N} = 0.55 \pm z_c \sqrt{0.55 \times 0.45/N} = 0.55 \pm 0.50 z_c / \sqrt{N}$, 同习题 9.10, 其中用到 $P = p = 0.55$. 因为只有候选人获得了 50% 以上的选票才能当选, 所以必须要使 $0.50 z_c / \sqrt{N}$ 小于 0.05.

(a) 对于 95% 的置信水平, $0.50 z_c / \sqrt{N} = 0.50 \times 1.96 / \sqrt{N} = 0.05$, $N = 384.2$. 则 N 至少为 385.

(b) 对于 99.73% 的置信水平, $0.50 z_c / \sqrt{N} = 0.50 \times 3 / \sqrt{N} = 0.05$, $N = 900$. 则 N 至少为 901.

另解 当 $\sqrt{N}/1.50 > 1/0.05$ 或 $\sqrt{N} > 1.50/0.05$ 时, $1.50/\sqrt{N} < 0.05$. 此时 $\sqrt{N} > 30$ 或 $N > 900$, 因此 N 至少为 901.

- 9.12 (a) 如果 P 是在容量为 N 的样本中观测到的成功比例. 证明: 在由 z_c 决定的置信水平上估计总体成功比例的置信限为

$$p = \frac{P + \frac{z_c^2}{2N} \pm z_c \sqrt{\frac{P(1-P)}{N} + \frac{z_c^2}{4N^2}}}{1 + \frac{z_c^2}{N}}$$

(b) 用 (a) 中得到的公式求习题 9.10 中的 99.73% 的置信限.

(c) 证明: 当 N 很大时, (a) 中的公式变为习题 9.10 中所用的 $p = P \pm z_c \sqrt{P(1-P)/N}$.

解 (a) 样本比例 P 的标准化变量为

$$\frac{P - p}{\sigma_P} = \frac{P - p}{\sqrt{p(1-p)/N}}$$

这一标准化变量的最大值和最小值是 $\pm z_c$, 其中 z_c 决定了置信水平. 让上式左端等于 $\pm z_c$, 我们有

$$P - p = \pm z_c \sqrt{\frac{p(1-p)}{N}}$$

将两边平方

$$P^2 - 2pP + p^2 = \frac{z_c^2 p(1-p)}{N}$$

两边同乘以 N 并化简, 得

$$(N + z_c^2)p^2 - (2NP + z_c^2)p + NP^2 = 0$$

令 $a = N + z_c^2$, $b = -(2NP + z_c^2)$ 且 $c = NP^2$, 这个方程变为 $ap^2 + bp + c = 0$, 将 p 看作未知数, 它的解为

$$p = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{2NP + z_c^2 \pm \sqrt{(2NP + z_c^2)^2 - 4(N + z_c^2) \cdot NP^2}}{2(N + z_c^2)}$$

$$= \frac{2NP + z_c^2 \pm z_c \sqrt{4NP(1-P) + z_c^2}}{2(N + z_c^2)}$$

分子和分母同除以 $2N$, 变为

$$p = \frac{P + \frac{z_c^2}{2N} \pm z_c \sqrt{\frac{P(1-P)}{N} + \frac{z_c^2}{4N^2}}}{1 + \frac{z_c^2}{N}}$$

(b) 对于 99.73% 的置信限, $z_c = 3$. 将 $P = 0.55$ 和 $N = 100$ 代入(a)中的公式, 得 $p = 0.40$ 和 0.69 , 和习题 9.10(a) 中的结果一致.

(c) 如果 N 很大, 那么 $z_c^2/(2N)$, $z_c^2/(4N^2)$ 和 z_c^2/N 都很小, 可忽略不计, 因而也可用零代替. 由此就可得要求的结果.

- 9.13 抛掷一枚硬币 40 次, 有 24 次出现正面. 无限次抛掷硬币, 求正面出现比例的(a) 95%, (b) 99.73% 的置信限.

解 9.13 (a) 在 95% 的置信水平上, $z_c = 1.96$. 将 $P = 24/40 = 0.6$ 和 $N = 40$ 代入习题 9.12(a) 中的公式, 得 $p = 0.45$ 和 0.74 . 因此我们可有 95% 的把握认为 p 在 0.45 和 0.74 之间.

由近似公式 $p = P \pm z_c \sqrt{P(1-P)/N}$ 得 $p = 0.60 \pm 0.15$, 由此得到的区间为 0.45 到 0.75 .

(b) 在 99.73% 的水平上, $z_c = 3$. 由习题 9.12(a) 中的公式, 得 $p = 0.37$ 和 0.79 .

由近似公式 $p = P \pm z_c \sqrt{P(1-P)/N}$ 得 $p = 0.60 \pm 0.23$, 由此得到的区间为 0.37 到 0.83 .

差与和的置信区间

- 9.14 150 个 A 品牌灯泡的平均寿命为 1400 小时, 标准差为 120 小时. 200 个 B 品牌灯泡的平均寿命为 1200 小时, 标准差为 80 小时. 求 A 和 B 两种品牌灯泡平均寿命之差的(a) 95%, (b) 99% 的置信限.

解 9.14 A 和 B 两种品牌灯泡平均寿命之差的置信限为

$$\bar{X}_A - \bar{X}_B \pm z_c \sqrt{\sigma_A^2/N_A + \sigma_B^2/N_B}$$

(a) 95% 的置信限是 $1400 - 1200 \pm 1.96 \sqrt{120^2/150 + 80^2/100} = 200 \pm 24.8$. 因此我们可以有 95% 的把握断定总体均值之差在 175 小时和 225 小时之间.

(b) 99% 的置信限是 $1400 - 1200 \pm 2.58 \sqrt{120^2/150 + 80^2/100} = 200 \pm 32.6$. 因此我们可以有 99% 的把握断定总体均值之差在 167 小时和 233 小时之间.

- 9.15 从收看某电视节目的观众中随机选择了 400 名成年人和 600 名青年人作调查. 其中有 100 名成年人和 300 名青年人表示喜爱此节目. 求在所有收看此节目的观众中, 成年人和青年人喜爱此节目的人数比例之差的(a) 95%, (b) 99% 的置信限.

解 9.15 两组人群中比例之差的置信限为

$$P_1 - P_2 \pm z_c \sqrt{p_1 q_1 / N_1 + p_2 q_2 / N_2}$$

其中下标 1 和 2 分别指青年人和成年人. $P_1 = 300/600 = 0.50$ 和 $P_2 = 100/400 = 0.25$ 分别为青年人和成年人中喜爱此节目的人数的比例.

(a) 95% 的置信限为 $0.50 - 0.25 \pm 1.96 \sqrt{0.50 \times 0.50/600 + 0.25 \times 0.75/400} = 0.25 \pm 0.06$. 因此我们可以有 95% 的把握断定真实的比例之差在 0.19 和 0.31 之间.

(b) 99% 的置信限为 $0.50 - 0.25 \pm 2.58 \sqrt{0.50 \times 0.50/600 + 0.25 \times 0.75/400} = 0.25 \pm 0.08$. 因此我们可以有 99% 的把握断定真实的比例之差在 0.17 和 0.33 之间.

- 9.16 某公司生产的电池的平均电压为 45.1 伏特, 标准差为 0.04 伏特. 将四节这样的电池串联起来, 求总电压的(a) 95%, (b) 99%, (c) 99.73%, (d) 50% 的置信限.

解 9.16 如果用 E_1, E_2, E_3 和 E_4 表示四节电池的电压, 则有

$$\begin{aligned} \mu_{E_1+E_2+E_3+E_4} &= \mu_{E_1} + \mu_{E_2} + \mu_{E_3} + \mu_{E_4} \\ \sigma_{E_1+E_2+E_3+E_4} &= \sqrt{\sigma_{E_1}^2 + \sigma_{E_2}^2 + \sigma_{E_3}^2 + \sigma_{E_4}^2} \end{aligned}$$

因为 $\mu_{E_1} = \mu_{E_2} = \mu_{E_3} = \mu_{E_4} = 45.1$ 且 $\sigma_{E_1} = \sigma_{E_2} = \sigma_{E_3} = \sigma_{E_4} = 0.04$, 则有 $\mu_{E_1+E_2+E_3+E_4} = 4 \times 45.1 = 180.4$ 和 $\sigma_{E_1+E_2+E_3+E_4} = \sqrt{4 \times 0.04^2} = 0.08$.

(a) 95% 的置信限为 $180.4 \pm 1.96 \times 0.08 = 180.4 \pm 0.16$ 伏特.

(b) 99% 的置信限为 $180.4 \pm 2.58 \times 0.08 = 180.4 \pm 0.21$ 伏特.

(c) 99.73% 的置信限为 $180.4 \pm 3 \times 0.08 = 180.4 \pm 0.24$ 伏特.

(d) 50% 的置信限为 $180.4 \pm 0.6745 \times 0.08 = 180.4 \pm 0.054$ 伏特. 值 0.054 伏特叫做可能误差.

标准差的置信区间

- 9.17 200 个灯泡样品平均寿命的标准差为 100 小时. 求所有灯泡的标准差的 (a) 95%, (b) 99% 的置信限.

解 总体标准差 σ 的置信限为 $s \pm z_c \sigma / \sqrt{2N}$, 其中 z_c 对应了置信水平. 用样本标准差估计 σ .

(a) 95% 的置信限是 $100 \pm 1.96 \times 100 / \sqrt{400} = 100 \pm 9.8$. 因此我们可以有 95% 的把握断定总体标准差在 90.2 小时和 109.8 小时之间.

(b) 99% 的置信限是 $100 \pm 2.58 \times 100 / \sqrt{400} = 100 \pm 12.9$. 因此我们可以有 99% 的把握断定总体标准差在 87.1 小时和 112.9 小时之间.

- 9.18 在习题 9.17 中, 为了有 99.73% 的把握确定总体标准差与样本标准差之差相对于样本标准差不超过 (a) 5%, (b) 10%, 必须要抽取多大的样本?

解 用 s 估计 σ , σ 的 99% 的置信限为 $s \pm 3\sigma / \sqrt{2N} = s \pm 3s / \sqrt{2N}$, 则总体标准差与样本标准差之差相对于样本标准差的百分比为

$$\frac{3s / \sqrt{2N}}{s} = \frac{300}{\sqrt{2N}} \%$$

(a) 如果 $300 / \sqrt{2N} = 5$, 则 $N = 1800$. 因此样本容量应为 1800 或更大.

(b) 如果 $300 / \sqrt{2N} = 10$, 则 $N = 450$. 因此样本容量应为 450 或更大.

可能误差

- 9.19 50 节同种型号电池的电压均值为 18.2 伏特, 标准差为 0.5 伏特. 求 (a) 均值的可能误差, (b) 50% 的置信限.

解 (a)

$$\text{均值的可能误差} = 0.6745 \sigma_{\bar{X}} = 0.6745 \frac{\sigma}{\sqrt{N}} = 0.6745 \frac{s}{\sqrt{N}}$$

$$= 0.6745 \frac{s}{\sqrt{N-1}} = 0.6745 \frac{0.5}{\sqrt{49}} = 0.048 \text{ 伏特}$$

注意: 如果用 $s = 0.5$ 代替 σ , 可能误差是 $0.6745 \times (0.5 / \sqrt{50}) = 0.048$. 所以若 N 足够大, 两种估计都可用.

(b) 50% 的置信限是 18 ± 0.048 伏特.

- 9.20 一个测量结果被记录为 216.480 克, 可能误差为 0.272 克. 这一测量结果的 95% 的置信限是多少?

解 可能误差是 $0.272 = 0.6745 \sigma_{\bar{X}}$ 或 $\sigma_{\bar{X}} = 0.272 / 0.6745$. 则 95% 的置信限是 $\bar{X} \pm 1.96 \sigma_{\bar{X}} = 216.480 \pm 1.96 \times (0.272 / 0.6745) = 216.480 \pm 0.790$ 克.

补充习题

无偏估计和有效估计

- 9.21 测量一组物体质量分别为 8.3, 10.6, 9.7, 8.8, 10.2 和 9.4 千克. 求 (a) 总体均值的无偏和有效估计, (b) 总体方差的无偏和有效估计, (c) 比较样本标准差和总体标准差.

- 9.22 一家公司生产的 10 个电子管样品的平均寿命为 1200 小时,标准差为 100 小时.估计这家公司生产的所有电子管寿命的 (a) 均值, (b) 标准差.
- 9.23 (a) 如果有 30, 50 和 100 个电子管样品, 求解习题 9.22.
(b) 对于不同容量的样本, 关于样本标准差和总体标准差的估计之间的关系, 你能得到什么结论?

均值的置信区间

- 9.24 60 条缆绳所承受的最大负荷的均值和标准差(见习题 3.59)分别是 11.09 吨和 0.73 吨. 求这家公司生产的所有缆绳的最大负荷均值的 (a) 95% 置信限, (b) 99% 置信限.
- 9.25 一家公司生产的 250 个铆钉顶端直径的均值和标准差分别为 0.72642 英寸和 0.00058 英寸(见习题 3.61). 求这家公司生产的所有铆钉顶端的平均直径的 (a) 99%, (b) 98%, (c) 95%, (d) 90% 的置信限.
- 9.26 求习题 9.25 中平均直径的 (a) 50% 的置信限, (b) 可能的误差.
- 9.27 如果电子管寿命的标准差估计为 100 小时, 则要取多大的样本才能有 (a) 95%, (b) 90%, (c) 99%, (d) 99.73% 的把握保证平均寿命的估计误差不超过 20 小时?
- 9.28 在习题 9.27 中, 如果平均寿命的估计误差不超过 10 小时, 样本容量应有多少?
- 9.29 一家公司有 500 条电缆, 从中随机抽取 40 条检验, 拉断力的均值为 2400 磅, 标准差为 150 磅.
(a) 剩下的 460 条电缆的拉断力均值的 95% 和 99% 的置信限是多少?
(b) 我们有多大把握保证剩下的 460 条电缆的拉断力的均值在 2400 ± 35 磅之间?

比例的置信区间

- 9.30 一个罐子里装有比例未知的红球和白球. 从中有放回地抽取 60 个球, 其中有 70% 是红球. 求罐子中红球真正比例的 (a) 95%, (b) 99%, (c) 99.73% 置信限. 分别用习题 9.12 中的近似公式和更精确的公式求出结果.
- 9.31 在习题 9.30 中, 应抽取多大的样本才能有 (a) 95%, (b) 99%, (c) 99.73% 的把握保证真正的比例和样本比例之差相对于样本比例不多于 5%?
- 9.32 可以相信, 两位候选人之间将会获得非常接近的选票, 问至少应调查多少选民才能有 (a) 80%, (b) 90%, (c) 95%, (d) 99% 的把握确定应支持哪一位候选人?

差与和的置信区间

- 9.33 有 A 和 B 两组病情类似的病人, 分别有 50 人和 100 人. 第一组服用了一种新型的安眠药, 第二组则服用普通的安眠药. A 组病人睡眠时间的平均值为 7.82 小时, 标准差为 0.24 小时. B 组病人睡眠时间的平均值为 6.75 小时, 标准差为 0.30 小时. 求由这两种药导致的平均睡眠时间之差的 (a) 95%, (b) 99% 的置信限.
- 9.34 一台机器生产的 200 个螺栓样品中有 15 个次品, 另一台机器生产的 100 个螺栓样品中有 12 个次品. 求这两台机器生产的产品中次品比例之差的 (a) 95%, (b) 99%, (c) 99.73% 的置信限. 并讨论得到的结果.
- 9.35 一家公司生产的轴承滚珠的平均重量为 0.638 磅, 标准差为 0.012 磅. 求 100 个轴承滚珠重量的 (a) 95%, (b) 99% 的置信限.

标准差的置信区间

- 9.36 一家公司检验的 100 条电缆的拉断力的标准差为 180 磅, 求这家公司生产的所有电缆拉断力的标准差的 (a) 95%, (b) 99%, (c) 99.73% 的置信限.
- 9.37 求习题 9.36 中标准差的可能误差.
- 9.38 要抽取多大的样本才能有 (a) 95%, (b) 99%, (c) 99.73% 的把握保证总体标准差和样本标准差之差相对于样本标准差不多于 2%?

第十章 统计决策理论

统计决策

在实际问题中,我们常常需要根据样本的信息对总体情况作出决策.这些决策称为**统计决策**.例如,我们希望根据样本数据来判断一种新药在治疗某种疾病时是否真正有效,一种教学方法是否优于其他的方法,或是一枚给定的硬币是否负重.

统计假设

要作出某些决策,常常要对总体先作出某些假定(或猜测).这些假定可能正确也可能不正确,称为**统计假设**.它们一般是关于总体的概率分布的某些陈述.

原假设

在很多情况下,我们给出一个统计假设仅仅是为了拒绝它.例如,如果我们要判断给定的一枚硬币是否负重,则我们假设硬币是均匀的(即 $p = 0.5$, 其中 p 是正面出现的概率).类似地,如果我们要判断一种方法是否优于其他的方法,则我们假设两种方法之间没有**差异**(即观察到的差别仅仅是由同一总体中抽样的波动性引起的).这样的假设常称为**零假设**或**原假设**,记为 H_0 .

备择假设

任何不同于零假设的假设都称为**备择假设**.例如,如果零假设是 $p = 0.5$,则备择选择可以是 $p = 0.7$, $p \neq 0.5$ 或 $p > 0.5$.备择假设记为 H_1 .

假设检验,显著性检验或决策法则

如果我们假定某个假设是正确的,但又发现观测到的随机抽样的结果和假设下预期的结果有显著的差别,则我们可以说观测到的差别是显著的,并拒绝该假设(至少根据观测到的证据不能接受它).例如,抛掷一枚硬币 20 次有 16 次出现正面,即使我们相信自己有可能会犯错,我们也会拒绝硬币是均匀的这一假设.

使我们能够判断观测到的样本是否和预期的结果有显著的区别,并帮助我们决定是否接受或拒绝假设的过程叫做**假设检验**,**显著性检验**或**决策法则**.

第一类和第二类错误

当我们拒绝了一个本应接受的假设时,我们就犯了**第一类错误**.反之,当我们接受了一个本应拒绝的假设时,我们就犯了**第二类错误**.无论是哪种情形,在判断上我们都犯了错误,作出了错误的决策.

为了使决策法则(或假设检验)更好,就必须要使决策错误达到最小.这并不是一个简单的问题,因为对于任何给定的样本容量,在减少一种类型错误的同时往往会使另一种类型的错误增加.在实际问题中,犯一种类型的错误可能比犯另一种类型的错误更加严重,这时就要作出一些妥协来限制更为严重的那类错误.同时减少两种类型错误的惟一方法是增加样本容量,而这未必都能做到.

显著性水平

在作假设检验时,我们愿意犯第一类错误的最大概率称为检验的**显著性水平**.这个概率常

记为 α , 通常都在抽样前就指定好, 这样得到的结果才不会影响我们的选择.

在实际问题中, 显著性水平可以有多种选择, 但最为普通的是 0.05 或 0.01. 例如, 如果设计一个决策法则选择的显著性水平是 0.05 (或 5%), 那么在 100 次中可能有 5 次机会使我们拒绝本该接受的假设; 也就是说, 我们大约有 95% 的把握作出正确的决策. 此时, 我们说拒绝假设的显著性水平为 0.05, 即犯拒绝本应接受的假设这类错误的概率是 0.05.

关于正态分布的检验

为了说明前面提到的思想, 假定在一个给定的假设之下某个统计量 S 的抽样分布服从均值为 μ_S 、标准差为 σ_S 的正态分布. 则标准化变量 (或 z 分数) $z = (S - \mu_S) / \sigma_S$ 的分布是标准正态分布 (均值为 0, 方差为 1).

如图 10-1 所示, 我们能有 95% 的把握断定: 如果假设是正确的, 则一个实际的样本统计量 S 的 z 分数在 -1.96 和 1.96 之间 (因为在正态曲线下这些值之间的面积为 0.95). 但是, 如果随机选取一个样本, 我们可能会发现这个统计量的 z 分数在 $(-1.96, 1.96)$ 之外. 如果假设是正确的, 我们可以肯定这样的事件发生的概率只有 0.05 (图中阴影部分的总面积). 因此我们可以认为 z 分数和假设下的期望值之间有显著的差异, 从而会拒绝这一假设.

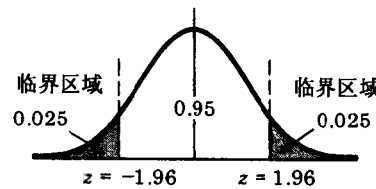


图 10-1

阴影部分的总面积 0.05 就是检验的显著性水平. 它表示了我们拒绝假设而犯错误的概率 (即犯第一类错误的概率). 所以我们可以说在显著性水平 0.05 下假设被拒绝或在水平 0.05 下给定的样本统计量的 z 分数是显著的.

在 $(-1.96, 1.96)$ 之外的 z 分数的集合构成了所谓的**假设的临界区域**, **假设的拒绝区域**或**显著性区域**. 在 $(-1.96, 1.96)$ 之内的 z 分数的集合称为**假设的接受区域**或**非显著区域**.

由上面的陈述, 我们可以总结出如下的决策法则 (或假设检验或显著性检验):

如果统计量 S 的 z 分数在 $(-1.96, 1.96)$ 之外 (即 $z > 1.96$ 或 $z < -1.96$), 则以 0.05 的显著性水平拒绝假设. 否则接受假设 (或如有需要, 不作任何决策).

因为 z 分数在假设检验中起到了重要的作用, 故也称之为**检验统计量**.

要注意的是也可采用其他显著性水平. 例如, 采用 0.01 的显著性水平, 我们就要将上面提到的 1.96 全部换成 2.58 (见表 10.1). 也可利用表 9.1, 因为显著性水平和置信水平之和为 100%.

双边检验和单边检验

在前面提到的检验中, 我们关心的是分布在均值两侧 (即分布的两侧) 的统计量 S 或相应的 z 分数的端值. 这样的检验叫做**双边检验**.

但是, 我们常常可能只对均值一侧 (即分布的一侧) 的端值感兴趣. 例如我们要检验一个程序优于其他程序这一假设 (不同于检验一个程序是否区别于其他程序). 这样的检验叫做**单边检验**. 此时, 临界区域位于分布的一侧, 它所对应的面积等于显著性水平.

表 10.1 给出了在不同的显著性水平下, z 在单边检验和双边检验中的临界值, 很具有参考价值. 对应于其他显著性水平的 z 的临界值可由正态曲线面积表 (附录 II) 求得.

表 10.1

显著性水平 α	0.10	0.05	0.01	0.005	0.002
z 在单边检验中的临界值	-1.28 或 1.28	-1.645 或 1.645	-2.33 或 2.33	-2.58 或 2.58	-2.88 或 2.88
z 在双边检验中的临界值	-1.645 和 1.645	-1.96 和 1.96	-2.58 和 2.58	-2.81 和 2.81	-3.08 和 3.08

特殊检验

在大样本情形,许多统计量的抽样分布都是正态分布(或至少接近于正态),相应的 z 分数就可应用于前面提到的检验.下面列出的特殊检验仅是实际中常用的几个检验,它们所用的统计量都取自表 8.1.对无限总体或从有限总体中有放回抽样,这些方法都适用.对从有限总体中无放回抽样,结论必须作些修正.

1. 均值:此时 $S = \bar{X}$ 为样本均值; $\mu_S = \mu_{\bar{X}} = \mu$ 为总体均值; $\sigma_S = \sigma_{\bar{X}} = \sigma / \sqrt{N}$,其中 σ 是总体标准差, N 是样本容量.则 z 分数为

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}}$$

如有必要,可用样本标准差 s 或 s 估计 σ .

2. 比例:此时 $S = P$ 为一个样本中“成功”的比例; $\mu_S = \mu_P = p$,其中 p 是总体中成功的比例; $\sigma_S = \sigma_P = \sqrt{pq/N}$,其中 $q = 1 - p$, N 是样本容量. z 分数为

$$z = \frac{P - p}{\sqrt{pq/N}}$$

如果 $P = X/N$,其中 X 是样本中成功的次数, z 分数变为

$$z = \frac{X - Np}{\sqrt{Npq}}$$

即 $\mu_X = \mu = Np$, $\sigma_X = \sigma = \sqrt{Npq}$ 且 $S = X$.对于其他统计量的结果可类似地得到.

OC 曲线,检验的功效

我们已经看到如何选择合适的显著性水平以限制第一类错误.而完全避免冒险犯第二类错误的惟一方法是不接受任何假设.但是在许多情况下,这都是不可能的.此时,常常要用到 **OC 曲线**(Operating-Characteristic Curve),这种曲线表明了在各种假设下犯第二类错误的概率,为如何在一个给定的检验中减少第二类错误提供了指示;也就是说,这种曲线表示了防止我们作出错误决策的**检验功效**.这在设计时是非常有用的,因为它常常能为我们指明所需的样本的容量.

控制图

在实际问题中,很重要的一件事是了解一个过程何时有了改变,从而需要采取一些弥补的措施.比如在质量控制中就会遇到这种问题.质量控制管理者常常要判定观测到的质量变化是偶尔的波动还是由生产过程中的一些真正变化引起的,如机器零件的损坏,员工的过失等.**控制图**为解决这类问题提供了一个简单有用的方法.

有关样本差的检验

均值之差

有两个均值分别为 μ_1 和 μ_2 ,标准差分别为 σ_1 和 σ_2 的总体,分别从中抽取一个样本容量为 N_1 和 N_2 (N_1, N_2 均较大)的样本,样本均值为 \bar{X}_1 和 \bar{X}_2 .考察原假设:两个总体均值之间没有**差异**(即 $\mu_1 = \mu_2$).也就是说样本是从两个均值相同的总体中抽取的.

将 $\mu_1 = \mu_2$ 代入第八章的(5)式,可得均值之差的抽样分布近似于正态分布,它的均值和方差为

$$\mu_{\bar{X}_1 - \bar{X}_2} = 0 \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \quad (1)$$

如有需要,可用样本标准差 s_1 和 s_2 (或 s_1 和 s_2) 估计 σ_1 和 σ_2 .

应用标准化变量

$$z = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} \quad (2)$$

我们可在一个适宜的显著性水平下对原假设与备择假设(或观测到的差异的显著性)作出检验.

比例之差

有两个总体的比例分别为 p_1 和 p_2 , 分别从中抽取一个样本容量为 N_1 和 N_2 (N_1, N_2 均较大) 的样本, 样本比例分别为 P_1 和 P_2 . 考察原假设: 两个总体参数之间没有差异(即 $p_1 = p_2$), 即样本是从相同的总体中抽取的.

将 $p_1 = p_2 = p$ 代入第八章(6)式, 可得比例之差的抽样分布近似于正态分布, 它的均值和方差分别为

$$\mu_{P_1 - P_2} = 0 \quad \sigma_{P_1 - P_2} = \sqrt{pq \left(\frac{1}{N_1} + \frac{1}{N_2} \right)} \quad (3)$$

其中 $p = \frac{N_1 P_1 + N_2 P_2}{N_1 + N_2}$ 是总体比例的一个估计, $q = 1 - p$.

应用标准化变量

$$z = \frac{P_1 - P_2 - 0}{\sigma_{P_1 - P_2}} = \frac{P_1 - P_2}{\sigma_{P_1 - P_2}} \quad (4)$$

我们可在一个适宜的显著性水平下检验原假设.

有关其他统计量的检验也可类似地设计.

关于二项分布的检验

关于二项分布(以及其他分布)的检验, 可通过用与正态分布的检验相类似的方法而得, 最基本的原理在本质上都是相同的. 见习题 10.23 至 10.28.

习题及解答

用正态分布检验均值和比例

10.1 抛掷一枚均匀硬币 100 次, 求有 40 次到 60 次出现正面的概率.

解 由二项分布, 要求的概率是

$$\binom{100}{40} \left(\frac{1}{2} \right)^{40} \left(\frac{1}{2} \right)^{60} + \binom{100}{41} \left(\frac{1}{2} \right)^{41} \left(\frac{1}{2} \right)^{59} + \cdots + \binom{100}{60} \left(\frac{1}{2} \right)^{60} \left(\frac{1}{2} \right)^{40}$$

因为 $Np = 100 \times \frac{1}{2}$ 和 $Nq = 100 \times \frac{1}{2}$ 都比 5 大, 因此可用二项分布的正态逼近来求这个和. 抛掷硬币 100 次, 正面出现次数的均值和标准差为

$$\mu = Np = 100 \times \frac{1}{2} = 50 \quad \sigma = \sqrt{Npq} = \sqrt{100 \times \frac{1}{2} \times \frac{1}{2}} = 5$$

从连续的角度来考虑, 40 到 60 次出现正面相当于 39.5 到 60.5 次出现正面. 则有

$$39.5 \text{ 的标准值} = \frac{39.5 - 50}{5} = -2.10$$

$$60.5 \text{ 的标准值} = \frac{60.5 - 50}{5} = 2.10$$

要求的概率 = (正态曲线下 $z = -2.10$ 和 $z = 2.10$ 之间的面积)

$$= 2 \times (z = 0 \text{ 和 } z = 2.10 \text{ 之间的面积}) = 2 \times 0.4821 = 0.9642$$

10.2 为了检验一枚硬币是均匀的, 采用如下的决策法则:

如果抛掷硬币 100 次有 40 到 60 次出现正面, 则接受假设. 否则拒绝假设.

(a) 求拒绝了正确假设的概率.

(b) 作图表示决策法则和(a)中的结果.

(c) 如果抛掷硬币 100 次有 53 次出现正面,你能得到什么结论?若有 60 次出现正面,又能得到什么结论?

(d) 你在(c)中得到的结论是否有可能出错?并解释原因.

解 (a) 由习题 10.1, 如果硬币是均匀的, 则出现正面的次数不在 40 到 60 之间的概率为 $1 - 0.9642 = 0.0358$. 则拒绝了正确假设的概率是 0.0358.

(b) 图 10-2 表示的是决策法则, 给出了抛掷均匀硬币 100 次正面出现次数的概率分布. 若由包含 100 次抛掷的样本得到的 z 分数在 -2.10 和 2.10 之间, 我们就接受假设; 否则就拒绝假设, 断定硬币是不均匀的.

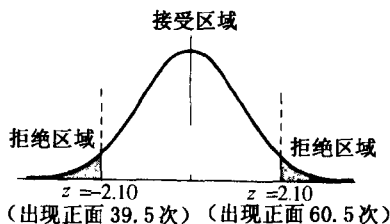


图 10-2

拒绝了本应接受的假设犯了**第一类错误**. 由(a), 犯这一错误的概率等于 0.0358, 即为图中阴影部分所示. 如果由包含 100 次抛掷的样本得到的 z 分数(或 z 统计量)在阴影区域内, 我们可以认为 z 分数与在假设假设是正确的条件下得到的结果有显著的不同. 为此, 阴影部分的总面积(即第一类错误的概率)称为决策法则的**显著性水平**, 在本题中显著性水平等于 0.0358. 因此, 我们可以说在 0.0358(或 3.58%) 的显著性水平下拒绝了假设.

(c) 根据决策法则, 在两种情形下我们都不得不接受硬币是均匀的这一假设. 有人会争辩说只要再多一次出现正面, 我们就要拒绝假设. 而这是我们在制定决策法则划定界限时必须要面临的问题.

(d) 有可能. 我们可能会接受了本应拒绝的假设. 如在本题中, 正面出现的概率也许是 0.7 而不是 0.5. 接受了本应拒绝的假设犯了**第二类错误**(更详细的讨论见习题 10.10 至 10.12).

10.3 抛掷一枚硬币 64 次, 在(a) 0.05, (b) 0.01 的显著性水平下, 制定一个决策法则检验硬币的均匀性.

解 (a) **解法一** 如果显著性水平是 0.05, 则由对称性, 图 10-3 中每一边的面积都是 0.025. 因此, 0 和 z_1 之间的面积为 $0.5000 - 0.0250 = 0.4750$, $z_1 = 1.96$. 也可从表 10.1 中得到临界值 -1.96 和 1.96 . 所以一个可能的决策法则:

如果 z 在 -1.96 和 1.96 之间, 则接受硬币是均匀的这一假设. 否则就拒绝假设.

为了用 64 次抛掷中正面出现的次数表达决策法则, 注意到如果假设是正确的, 正面出现次数的分布的均值和标准差为

$$\mu = Np = 64 \times 0.5 = 32 \quad \sigma = \sqrt{Npq} = \sqrt{64 \times 0.5 \times 0.5} = 4$$

则 $z = (X - \mu)/\sigma = (X - 32)/4$. 若 $z = 1.96$, 则 $(X - 32)/4 = 1.96$, $X = 39.84$; 若 $z = -1.96$, 则 $(X - 32)/4 = -1.96$, $X = 24.16$. 因此决策法则变为:

如果正面出现的次数在 24.16 和 39.84(即 25 和 39)之间, 则接受硬币是均匀的这一假设. 否则就拒绝假设.

解法二 正面出现的次数以 0.95 的概率在 $\mu - 1.96\sigma$ 和 $\mu + 1.96\sigma$ (即 $Np - 1.96\sqrt{Npq}$ 和 $Np + 1.96\sqrt{Npq}$ 之间), 或在 $32 - 1.96 \times 4 = 24.16$ 和 $32 + 1.96 \times 4 = 39.84$ 之间. 由此即可得上述的决策法则.

解法三 $-1.96 < z < 1.96$ 就相当于 $-1.96 < \frac{1}{4}(X - 32) < 1.96$, 则 $-1.96 \times 4 < X - 32 < 1.96 \times 4$, 也就是 $32 - 1.96 \times 4 < X < 32 + 1.96 \times 4$ (即 $24.16 < X < 39.84$), 这就得到了上述的决策法则.

(b) 如果显著性水平是 0.01, 图 10-3 中每一边的阴影部分面积就是 0.005. 则 0 和 z_1 之间的面积是 $0.5000 - 0.0050 = 0.4950$, $z_1 = 2.58$ (更精确为 2.575), 这也可从表 10.1 中找到. 同(a)中解法二的步骤, 正面出现的次数以 0.99 的概率

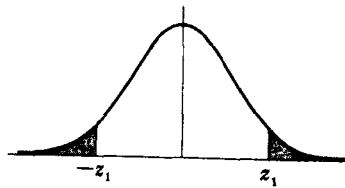


图 10-3

在 $\mu - 2.58\sigma$ 和 $\mu + 2.58\sigma$, 或 $32 - 2.58 \times 4 = 21.68$ 和 $32 + 2.58 \times 4 = 42.32$ 之间. 因此决策法则本

为:

如果正面出现的次数在 22 和 42 之间,则接受硬币是均匀的这一假设.否则就拒绝假设.

10.4 在习题 10.3 中,如何制定一个决策法则以避免犯第二类错误?

解 接受了本应拒绝的假设就犯了第二类错误.我们可以这样避免犯此类错误:以不拒绝假设代替接受假设.这就意味着在这种情形下我们并没有给出任何决策.例如,我们可将习题 10.3(b)中的决策法则改为:

如果正面出现的次数在 22 和 42 之间,则不拒绝硬币是均匀的这一假设.否则就拒绝假设.

但是在很多实际问题中,决定接受或拒绝假设非常重要.这种情形下的详细讨论要考虑到第二类错误(见习题 10.10 至 10.12).

10.5 在一个关于超感觉(ESP)力的试验中,一间房间里的一个人从洗好的 50 张卡片中随机抽取一张,而在另一间房间里的人(主体)要说出卡片的颜色(红或蓝).接受试验的主体并不知道有多少张红色和蓝色的卡片.如果主体能准确地说出 32 张卡片的颜色,在(a) 0.05, (b) 0.01 的水平下判定结果是否是显著的.

解 如果 p 表示主体能准确地说出卡片颜色的概率,则我们不得不在下面的两个假设中作出选择:

$H_0: p = 0.5$, 主体只是单凭猜测(即结果只是出于偶然).

$H_1: p > 0.5$, 主体有 ESP 力.

因为我们并不关心得分低的主体,而只是关心那些得分高的主体,所以选择用单边检验.假定 H_0 是正确的,则能被准确地说出颜色的卡片张数的均值和标准差为

$$\mu = Np = 50 \times 0.5 = 25 \quad \sigma = \sqrt{Npq} = \sqrt{50 \times 0.5 \times 0.5} = \sqrt{12.5} = 3.54$$

(a) 对于显著性水平为 0.05 的单边检验,我们必须如图 10-4 选择 z_1 值,这样阴影部分表示的高分临界区域的面积为 0.05. 则 0 和 z_1 之间的面积是 0.4500, $z_1 = 1.645$, 这也可从表 10.1 中找到.所以,我们的决策法则(或显著性检验)为:

如果观测的 z 分数大于 1.645, 则结果在 0.05 的水平下是显著的,主体有 ESP 力.如果 z 分数小于 1.645, 则结果只是出于偶然(即在 0.05 的水平下不显著).

因为 32 的标准值是 $(32 - 25)/3.54 = 1.98$, 大于 1.645, 我们可以断定在 0.05 的水平下,主体有 ESP 力.

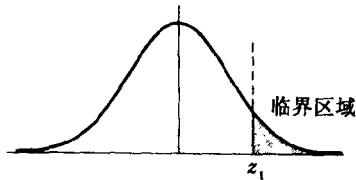


图 10-4

注意,实际上,我们应该作一个关于连续的修正.而从连续的角度来考虑,32 在 31.5 和 32.5 之间.31.5 的标准值是 $(31.5 - 25)/3.54 = 1.84$, 所以可得到相同的结论.

(b) 如果显著性水平是 0.01, 0 和 z_1 之间的面积是 0.4900, $z_1 = 2.33$. 由于 32(或 31.5)的标准值是 1.98(或 1.84), 小于 2.33, 我们可以断定结果在 0.01 的水平下不显著.

有些统计学家这样采用术语表达:如果结果在 0.01 的水平下显著,则称为**高度显著**.如果在 0.05 的水平下显著,但在 0.01 的水平下不显著则称为**可能显著**.如果在大于 0.05 的水平下显著,则称为**不显著**.根据这些术语,我们可以断定上述的试验结果可能显著,所以还需要就此再做一些深入的研究.

因为在作决策时,显著性水平相当于一个向导,所以有些统计学家就引用了所谓的真实概率.例如在本题中, $P(z \geq 1.84) = 0.0322$, 统计学家会说:根据这个试验,100 次中大约会有 3 次机会错误判定某个人有 ESP 力.被引用的概率(本题中是 0.0322)称为检验的 p 值.

10.6 某成药制造商声明他有 90% 的把握保证他的药能有效缓解过敏达 8 小时.从患有过敏症的人中随机抽取 200 人,服药后有 160 人的症状得到了缓解.判定此制造商的声明是否真实.

解 用 p 表示服药后过敏症得到缓解的概率.则我们必须在下面的两个假设中作出选择:

$H_0: p = 0.9$, 声明是真实的.

$H_1: p < 0.9$, 声明是不真实的.

因为我们只关心得到缓解的比例是否太低,所以可选择单边检验.如果显著性水平是 0.01(即如果图

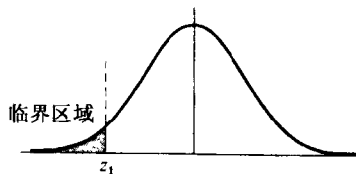


图 10-5

10-5 中阴影部分的面积是 0.01), 则可同习题 10.5(b) 一样利用曲线的对称性或查表 10.1 得 $z_1 = -2.33$. 因此我们的决策法则为:

如果 z 小于 -2.33 , 则声明是不真实的 (此时我们拒绝假设 H_0). 否则声明是真实的, 观测到的结果只是出于偶然 (此时我们接受假设 H_0).

若 H_0 是真的, 则 $\mu = Np = 200 \times 0.9 = 180$ 且 $\sigma = \sqrt{Npq} = \sqrt{200 \times 0.9 \times 0.1} = 4.24$. 故 160 的标准值是

$(160 - 180)/4.24 = -4.72$, 于 -2.33 . 所以, 根据我们的决策法则可断定声明是不真实的且试验结果是高度显著的 (见习题 10.5 的结尾).

- 10.7 我们规定一个假设检验的 p -值是拒绝零假设的最小显著性水平. 本题说明了计算检验统计量 p 值的方法. 用习题 9.6 中的数据检验零假设: 所有树木的平均高度等于 5 英尺. 备择假设: 平均高度小于 5 英尺. 找出这个检验的 p -值.

解 计算出的 z 值是 $z = (59.22 - 60)/1.01 = -0.77$. 拒绝零假设的最小显著性水平就是 p 值, 为 $P(z < -0.77) = 0.5 - 0.2794 = 0.2206$. 如果 p 小于预先给定的显著性水平, 就要拒绝零假设. 在本题中, 如果显著性水平给定为 0.05, 零假设就不会被拒绝. Minitab 的结果如下所示, 其中子命令 Alternative - 1 表示细尾检验.

MTB> ZTest mean = 60 sd = 10.111 data in cl;

SUBC> Alternative - 1.

Z-Test

Test of $\mu = 60.00$ vs $\mu < 60.00$

The assumed sigma = 10.1

Variable	N	Mean	StDev	SE Mean	Z	P
height	100	59.22	10.11	1.01	-0.77	0.22

- 10.8 从收听收音机的人群中随机抽取 33 个人, 记录下他们每周收听的时间, 以小时为单位的数据如下:

9 8 7 4 8 6 8 8 7 10 8 10 6 7 7 8 9
6 5 8 5 6 8 7 8 5 5 8 7 6 6 4 5

在 $\alpha = 0.05$ 的显著性水平下, 以下列三种方式检验 $\mu = 5$ 的零假设对 $\mu \neq 5$ 的备择假设:

- 计算检验统计量的值, 并和 $\alpha = 0.05$ 的临界值相比较.
- 计算对应于已求出的检验统计量的 p 值, 并将 p 值和 $\alpha = 0.05$ 相比较.
- 求 μ 的 $1 - \alpha = 0.95$ 的置信区间, 并判断 5 是否落在此区间内.

解 在下面的 Minitab 的输出结果中, 先给出了标准差, 然后又分别给出了 Ztest 和 Zinterval 的结果.

MTB> standard deviation cl

Standard deviation of hours = 1.6005

MTB> ZTest 5.0 1.6005 'hours';

SUBC> Alternative 0.

Z-Test

Test of $\mu = 5.000$ vs $\mu \text{ not } = 5.000$

The assumed sigma = 1.60

Variable	N	Mean	StDev	SE Mean	z	p
hours	33	6.897	1.600	0.279	6.81	0.0000

MTB> ZInterval 95.0 1.6005 'hours'.

Variable	N	Mean	StDev	SE Mean	95.0% CI
hours	33	6.897	1.600	0.279	(6.351, 7.443)

(a) 计算得检验统计量的值为 $z = \frac{6.897 - 5}{0.279} = 6.81$, 临界值为 ± 1.96 , 因此要拒绝零假设. 注意, 计算出的值在 Minitab 的结果中已有显示.

(b) 由 Minitab 的结果 p 值为 0.0000, 因为 p 值 $< \alpha = 0.05$, 所以要拒绝零假设.

(c) 因为零假设指定的 5 并没包含在 μ 的 95% 的置信区间内, 所以要拒绝零假设.

这三种程序在检验零假设对双边的备择假设时是等价的.

10.9 某工厂生产的电缆的平均拉断力为 1800 磅, 标准差为 100 磅. 据说进行了技术革新后, 电缆的拉断力可以得到提高. 为了检验这一说法, 随机抽取 50 条电缆发现它们的平均拉断力是 1850 磅. 我们能否在 0.01 的显著性水平下支持这一说法?

解 我们必须下面两个假设中作出选择:

$H_0: \mu = 1800$ 磅, 拉断力并没有真正改变.

$H_1: \mu > 1800$ 磅, 拉断力确有改变.

这里要用单边检验, 与此相关的图和习题 10.5(a) 中的图 10-4 一样. 在 0.01 的显著性水平下, 决策法则如下:

如果观测到的 z 分数大于 2.33, 则结果在 0.01 的水平下是显著的, 要拒绝 H_0 . 否则就接受 H_0 (或不作出决策).

假定 H_0 是正确的, 我们有

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}} = \frac{1850 - 1800}{100 / \sqrt{50}} = 3.55$$

它大于 2.33. 因此, 我们可以断定结果是高度显著的且要支持这一说法.

OC 曲线

10.10 参习题 10.2, 当正面出现的真实概率 $p = 0.7$ 时, 接受硬币是均匀的概率是多少?

解 当抛掷硬币 100 次有 39.5 到 60.5 次出现正面时, 接受硬币是均匀 (即 $p = 0.5$) 的假设.

图 10-6 中左边的正态曲线下方阴影部分面积 α 表示了拒绝本应接受的假设 H_0 的概率 (即犯第一类错误的概率). 如习题 10.2(a) 中所求, 这一面积 α 表示的是检验的显著性水平, 等于 0.0358.

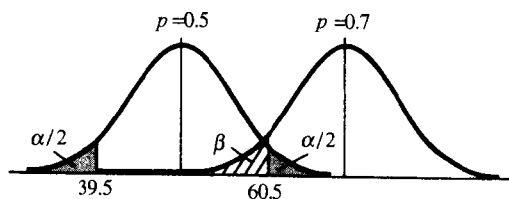


图 10-6

如果正面出现的概率是 $p = 0.7$, 那么 100 次中正面出现次数的分布由图 10-6 中右边的正态曲线所表示. 可以清楚地看到当 $p = 0.7$ 时接受 H_0 的概率 (即犯第二类错误的概率) 等于图中画斜线部分的面积 β . 为了计算这个面积, 我们求得在 $p = 0.7$ 的假设下分布的均值和标准差为

$$\mu = Np = 100 \times 0.7 = 70 \quad \sigma = \sqrt{Npq} = \sqrt{100 \times 0.7 \times 0.3} = 4.58$$

$$60.5 \text{ 的标准值} = \frac{60.5 - 70}{4.58} = -2.07 \quad 39.5 \text{ 的标准值} = \frac{39.5 - 70}{4.58} = -6.66$$

则

$$\beta = (\text{正态曲线下 } z = -6.66 \text{ 和 } z = -2.07 \text{ 之间的面积}) = 0.0192$$

因此根据给定的决策法则, 当 $p = 0.7$ 时接受硬币是均匀的机会很小.

注意, 在本题中, 已知决策法则, 再计算 α 和 β . 事实上, 还可能是下面两种情况:

(1) 由 α (如 0.05 或 0.01) 得到决策法则, 从而再计算 β .

(2) 由 α 和 β 得到决策法则.

10.11 当 (a) $p = 0.6$, (b) $p = 0.8$, (c) $p = 0.9$, (d) $p = 0.4$ 时求解习题 10.10.

解 (a) 若 $p=0.6$, 正面出现次数的分布的均值和标准差为

$$\begin{aligned}\mu &= Np = 100 \times 0.6 = 60 & \sigma &= \sqrt{Npq} = \sqrt{100 \times 0.6 \times 0.4} = 4.90 \\ 60.5 \text{ 的标准值} &= \frac{60.5 - 60}{4.90} = 0.102 \\ 39.5 \text{ 的标准值} &= \frac{39.5 - 60}{4.90} = -4.18\end{aligned}$$

则

$$\beta = (\text{正态曲线下 } z = -4.18 \text{ 和 } z = 0.102 \text{ 之间的面积}) = 0.5406$$

因此根据给定的决策法则, 当 $p=0.6$ 时接受硬币是均匀的机会很大.

(b) 若 $p=0.8$, 则

$$\begin{aligned}\mu &= Np = 100 \times 0.8 = 80 & \sigma &= \sqrt{Npq} = \sqrt{100 \times 0.8 \times 0.2} = 4 \\ 60.5 \text{ 的标准值} &= \frac{60.5 - 80}{4} = -4.88 \\ 39.5 \text{ 的标准值} &= \frac{39.5 - 80}{4} = -10.12\end{aligned}$$

则

$$\beta = (\text{正态曲线下 } z = -10.12 \text{ 和 } z = -4.88 \text{ 之间的面积}), \text{ 近似为 } 0.0000$$

(c) 和(b)相比较或通过计算知: 如 $p=0.9$, 则在实际问题中近似有 $\beta=0$.

(d) 由对称性, $p=0.4$ 和 $p=0.6$ 得到的 β 值相同 (即 $\beta=0.5040$).

10.12 作表示习题 10.10 和 10.11 中结果的(a) β 对 p , (b) $1-\beta$ 对 p 图, 并作出说明.

解 表 10.2 给出了习题 10.10 和 10.11 中得到的对应于 p 值的 β 值. 注意, β 表示的是当真正的 p 不等于 0.5 但接受了 $p=0.5$ 的概率; 若真正的 p 等于 0.5, 则 β 表示的是接受本应接受的 $p=0.5$ 的概率. 这一概率是 $1 - 0.0358 = 0.9642$, 也已列在了表 10.2 中.

表 10.2

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
β	0.0000	0.0000	0.0192	0.5040	0.9642	0.5040	0.0192	0.0000	0.0000

(a) β 对 p 的关系如图 10-7(a)所示, 这种图形叫做决策法则(或假设检验)的 OC 曲线. 曲线上的最高点和直线 $\beta=1$ 的距离等于 $\alpha=0.0358$, 即检验的显著性水平.

一般说来, OC 曲线的峰越尖, 决策法则在拒绝不成立的假设时效果越好.

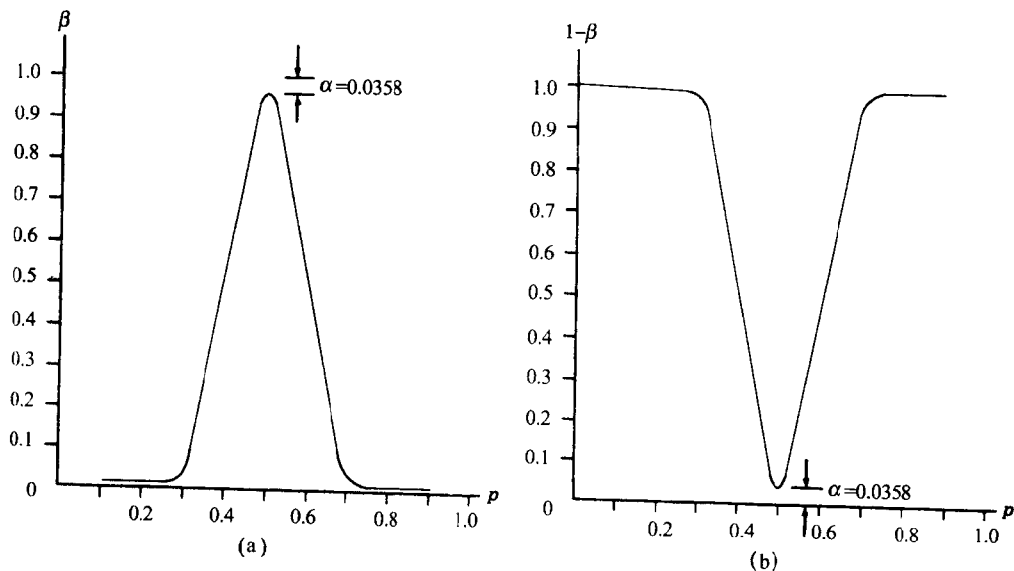


图 10-7

(b) $1 - \beta$ 对 p 的关系如图 10-7(b) 所示, 这种图形叫做决策法则的**功效曲线**, 只要将 OC 曲线倒置即得, 因此两种曲线实际上是等价的.

因为 $1 - \beta$ 表示的是拒绝本应拒绝的(或不成立的或错误的)假设(拒绝这种假设是应该的, 正确的)的**检验功效**, 所以常称为**功效函数**. β 也称为检验的**运算特性函数**.

10.13 某工厂生产的绳子的平均拉断力是 300 磅, 标准差是 24 磅. 现改进生产方法, 绳子的拉断力将会得到提高.

(a) 检验 64 条绳子, 在 0.01 的显著性水平下设计一个决策法则以废除旧的生产方法.

(b) 在(a)中得到的决策法则下, 当新的生产方法使得绳子的平均拉断力提高到 310 磅时, 检验结果却接受旧的方法的概率是多少? 假定标准差还是 24 磅.

解 (a) 若 μ 表示平均拉断力, 我们就要在下面的两个假设中作出选择:

$H_0: \mu = 300$ 磅, 新的方法和旧的一样.

$H_1: \mu > 300$ 磅, 新的方法比旧的好.

对于显著性水平为 0.01 的单边检验, 我们有如下的决策法则(参考图 10-8(a)):

如果样本平均拉断力的 z 分数大于 2.33 就拒绝 H_0 . 否则就接受 H_0 .

因为

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}} = \frac{\bar{X} - 300}{24 / \sqrt{64}}$$

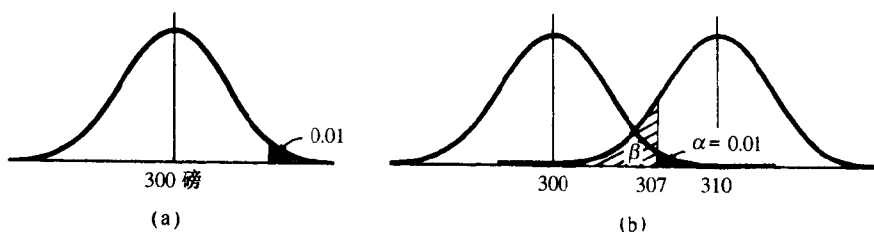


图 10-8

则有 $\bar{X} = 300 + 3z$. 若 $z > 2.33$, 则 $\bar{X} > 300 + 3 \times 2.33 = 307.0$ 磅. 因此, 上面的决策法则变为:

如果 64 条绳子的平均拉断力超过 307.0 磅就拒绝 H_0 . 否则就接受 H_0 .

(b) 考虑假设 $H_0: \mu = 300$ 磅和 $H_1: \mu = 310$ 磅. 图 10-8(b) 中的左右两条正态曲线分别表示这两个假设下的平均拉断力的分布. 图 10-8(b) 中的面积 β 表示的是当新的平均拉断力提高到 310 磅时, 检验结果却接受旧方法的概率. 为求得 β , 注意到 307.0 磅的标准值是 $(307.0 - 310)/3 = -1.00$, 因此

$$\beta = (\text{右边正态曲线下 } z = -1.00 \text{ 左方的面积}) = 0.1587$$

这就是当 $H_1: \mu = 310$ 磅成立时却接受 $H_0: \mu = 300$ 磅的概率(即犯第二类错误的概率).

10.14 假定拉断力的标准差仍是 24 磅, 作习题 10.13 中的(a) OC 曲线;(b) 功效曲线.

解 同习题 10.13(b), 当新的方法使得平均拉断力 μ 等于 305 磅, 315 磅时, 我们可求出相应的 β . 例如, 若 $\mu = 305$ 磅, 则 307.0 磅的标准值是 $(307.0 - 305)/3 = 0.67$, 因此

$$\beta = (\text{右边正态曲线下 } z = 0.67 \text{ 左方的面积}) = 0.7486$$

这样可得到表 10.3.

表 10.3

μ	290	295	300	305	310	315	320
β	1.0000	1.0000	0.9900	0.7486	0.1587	0.0038	0.0000

(a) OC 曲线如图 10-9(a) 所示. 由此曲线可见, 当新的平均拉断力小于 300 磅时保持旧的生产方法

的概率实际上是 1 (在新的平均值为 300 磅时是 1 减去显著性水平 0.01). 然后曲线就快速下降到零, 这表示当平均拉断力大于 315 磅时, 实际上没有任何可能保持旧的方法.

(b) 功效曲线如图 10-9(b) 所示. 对于功效曲线的解释和 OC 曲线一样. 事实上, 这两种曲线在本质上是等价的.

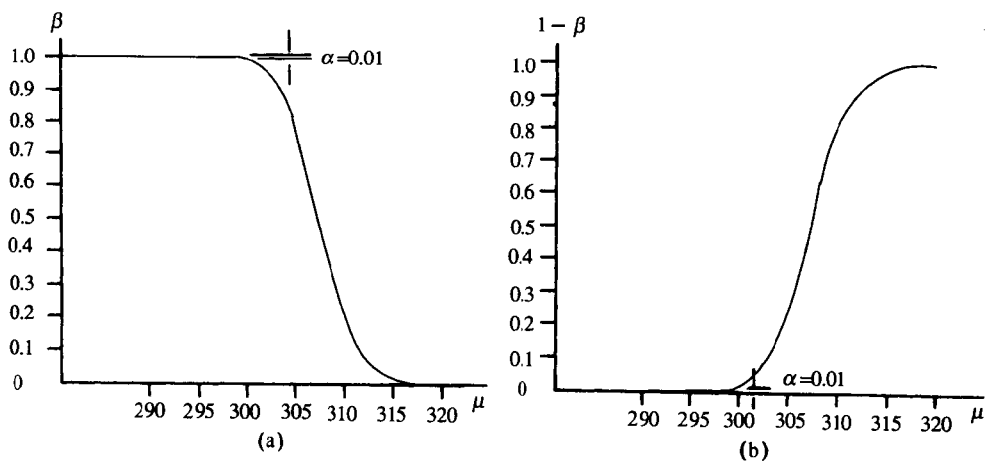


图 10-9

10.15 抛掷硬币若干次, 检验硬币是均匀 (即 $p = 0.5$) 的这一假设. 我们先作出如下的限制:

(1) 拒绝正确假设的概率至多为 0.05.

(2) 当真正的 p 和 0.5 之差的绝对值不小于 0.1 (即 $p \geq 0.6$ 或 $p \leq 0.4$) 时, 接受假设的概率至多为 0.05.

求所需样本容量的最小值, 并给出决策法则.

解 此题指定了冒险犯第一类和第二类错误的极限值. 例如, 限制(1)表示犯第一类错误的概率至多为 $\alpha = 0.05$, 而限制(2)表示犯第二类错误的概率至多为 $\beta = 0.05$, 如图 10-10 所示.

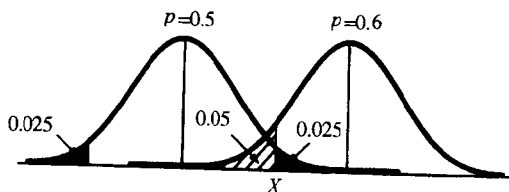


图 10-10

N 表示需要的样本容量, X 表示抛掷 N 次正面出现的次数. 如图 10-10, 我们要求 N, X 使得: $p = 0.5$ 时正态曲线在

$$\frac{X - Np}{\sqrt{Npq}} = \frac{X - 0.5N}{\sqrt{N \times 0.5 \times 0.5}} = \frac{X - 0.5N}{0.5\sqrt{N}} \quad (5)$$

右方的面积为 0.025, $p = 0.6$ 时正态曲线在

$$\frac{X - Np}{\sqrt{Npq}} = \frac{X - 0.6N}{\sqrt{N \times 0.6 \times 0.4}} = \frac{X - 0.6N}{0.49\sqrt{N}} \quad (6)$$

左方的面积为 0.05.

(事实上, 应使得 $(X - 0.6N)/(0.49\sqrt{N})$ 和 $[(N - X) - 0.6N]/(0.49\sqrt{N})$ 之间的面积是 0.05, (6) 式是一个很好的近似.) 由 (5) 式有

$$\frac{X - 0.5N}{0.5\sqrt{N}} = 1.96 \quad \text{或} \quad X = 0.5N + 0.980\sqrt{N} \quad (7)$$

由方程 (6)

$$\frac{X - 0.6N}{0.49\sqrt{N}} = -1.645 \quad \text{或} \quad X = 0.6N - 0.806\sqrt{N} \quad (8)$$

则由(7)和(8)式得 $N = 318.98$. 因此样本容量至少是 319 (即我们至少抛掷硬币 319 次). 将 $N = 319$ 代入(7)或(8), $X = 177$.

当 $p = 0.5$ 时, 有 $X - Np = 177 - 159.5 = 17.5$. 故采用如下的决策法则:

抛掷硬币 319 次, 如果正面出现的次数在 159.5 ± 17.5 之间 (即在 142 和 177 之间) 就接受假设. 否则就拒绝假设.

控制图

10.16 制造一台机器用以生产平均直径为 0.574 厘米, 标准差为 0.008 厘米的滚珠轴承. 为了检验该机器是否正常运作, 每隔 2 小时抽取一个含 6 个滚珠轴承的样本, 记录下样本的平均直径.

(a) 制定一个决策法则以确保产品质量符合要求.

(b) 说明如何作图表示(a)中得到的决策法则.

解 (a) 在 99.73% 的置信水平上, 样本均值 \bar{X} 在 $\mu_{\bar{X}} - 3\sigma_{\bar{X}}$ 和 $\mu_{\bar{X}} + 3\sigma_{\bar{X}}$ 之间或 $\mu - 3\sigma/\sqrt{N}$ 和 $\mu + 3\sigma/\sqrt{N}$ 之间. 因为 $\mu = 0.574$, $\sigma = 0.008$, $N = 6$, 所以在 99.73% 的置信水平上, 样本均值 \bar{X} 应在 $0.574 - 0.024/\sqrt{6}$ 和 $0.574 + 0.024/\sqrt{6}$ 之间或 0.564 和 0.584 之间. 因此得到下面的决策法则:

如果样本均值在 0.564 厘米和 0.584 厘米之间, 则机器运作正常. 否则, 机器就有问题需要检查.

(b) 将样本均值记录在均值图中, 如图 10-11, 称为质量控制图. 每抽取一个样本, 计算出它的均值, 并在图上用一个点表示. 只要点在下限(0.564 厘米)和上限(0.584 厘米)之间, 生产就在控制之中. 如果点跑出了控制限之外 (如星期四抽取的第三个样本), 那么就可能发生了什么问题, 必须作进一步的检查. 上面指定的控制限称为 99.73% 的置信限, 或简称为 3σ 限. 其他的置信限 (如 99% 或 95% 的置信限) 也可类似求得. 要根据不同的情况, 选择不同的置信限. 关于控制图的更详细的讨论见第十九章.

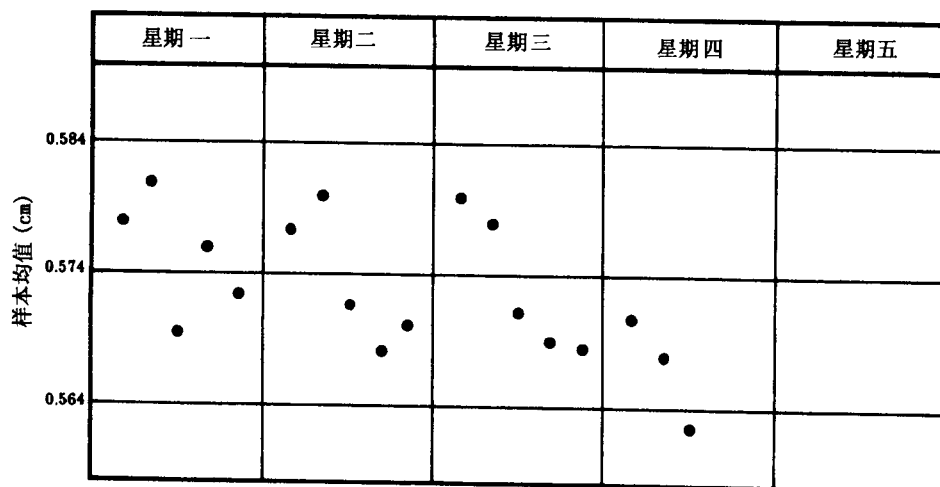


图 10-11

有关均值之差和比例之差的检验

10.17 在分别有 40 人和 50 人的两个班级进行测验. 其中, 一班的平均分是 74 分, 标准差是 8 分; 二班的平均分是 78 分, 标准差是 7 分. 在 (a) 0.05, (b) 0.01 的水平下两班成绩是否有显著区别?

解 假定两班学生分别来自于均值为 μ_1 和 μ_2 的两个总体. 我们要在下面的两个假设中作出选择:

$H_0: \mu_1 = \mu_2$, 区别只是出于偶然.

$H_1: \mu_1 \neq \mu_2$, 两者之间有显著区别.

在假设 H_0 下, 两班来自相同的总体. 样本均值之差的均值和标准差分别为

$$\mu_{\bar{X}_1 - \bar{X}_2} = 0 \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{8^2}{40} + \frac{7^2}{50}} = 1.606$$

其中, 我们用样本标准差作为 σ_1 和 σ_2 的估计. 则

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{74 - 78}{1.606} = -2.49$$

(a) 对于双边检验, 如果 z 在 -1.96 到 1.96 之外, 则结果在 0.05 的水平下是显著的. 因此我们可以断定在 0.05 的水平下, 两班成绩有显著差异, 二班成绩要好于一班.

(b) 对于双边检验, 如果 z 在 -2.58 到 2.58 之外, 则结果在 0.01 的水平下是显著的. 因此我们可以断定在 0.01 的水平下, 两班成绩没有显著差异.

因为结果在 0.05 的水平下显著, 但在 0.01 的水平下不显著, 所以我们可断定结果可能显著(根据习题 10.5 的结尾所用的术语).

- 10.18** 某大学参加运动会的 50 名男生的平均身高是 68.2 英寸, 标准差是 2.5 英寸. 其他 50 名男生的平均身高是 67.5 英寸, 标准差是 2.8 英寸. 检验假设: 参加运动会的男生的平均身高大于其他男生的平均身高.

解 我们必须下面的两个假设中作出选择:

$H_0: \mu_1 = \mu_2$, 两个平均身高之间没有区别.

$H_1: \mu_1 > \mu_2$, 第一组人的平均身高大于第二组人的平均身高.

在假设 H_0 下有

$$\mu_{\bar{X}_1 - \bar{X}_2} = 0 \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{2.5^2}{50} + \frac{2.8^2}{50}} = 0.53$$

其中, 我们用样本标准差估计 σ_1 和 σ_2 . 则

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{68.2 - 67.5}{0.53} = 1.32$$

用显著性水平为 0.05 的单边检验, 如果 z 分数大于 1.645 就拒绝 H_0 . 所以在这个水平上, 我们不能拒绝假设 H_0 .

但是, 要注意的是: 如果我们愿意冒险使犯第一类错误的概率是 0.10 (即 10 次中有 1 次机会), 则在 0.10 的水平上要拒绝假设 H_0 .

- 10.19** 在习题 10.18 中, 如果观测到的平均身高的差异是 0.7 英寸, 为了保证这个差异在 (a) 0.05 , (b) 0.01 的水平下是显著的, 两个样本的容量应该增加多少?

解 假设每一个样本的容量是 N . 在假设 H_0 下有

$$\mu_{\bar{X}_1 - \bar{X}_2} = 0 \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N} + \frac{\sigma_2^2}{N}} = \sqrt{\frac{2.5^2 + 2.8^2}{N}} = \sqrt{\frac{14.09}{N}} = \frac{3.75}{\sqrt{N}}$$

对于观测到的 0.7 英寸的差异, 我们有

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{0.7}{3.75/\sqrt{N}} = \frac{0.7\sqrt{N}}{3.75}$$

(a) 当 $0.7\sqrt{N}/3.75$ 至少为 1.645 时, 观测到的差异在 0.05 的水平下是显著的. 所以 N 至少是 78. 因此, 每一个样本的容量至少要增加 $78 - 50 = 28$.

另解

$$\frac{0.7\sqrt{N}}{3.75} \geq 1.645, \sqrt{N} \geq \frac{3.75 \times 1.645}{0.7}, \sqrt{N} \geq 8.8, N \geq 77.4 \text{ 或 } N \geq 78$$

(b) 如果

$$\frac{0.7\sqrt{N}}{3.75} \geq 2.33, \sqrt{N} \geq \frac{3.75 \times 2.33}{0.7}, \sqrt{N} \geq 12.5, N \geq 156.3 \text{ 或 } N \geq 157$$

则观测到的差异在 0.01 的水平下是显著的. 因此, 每一个样本的容量至少要增加 $157 - 50 = 107$.

- 10.20** A 组和 B 组各包含 100 个病人. A 组人使用了一种血清, 而 B 组人没有 (称为控制).

除 A 组人多使用该种血清外, 两组人都进行了同样的治疗, 结果 A 组和 B 组中各有 75 人和 65 人治愈. 在 (a) 0.01, (b) 0.05, (c) 0.10 的显著性水平上, 检验假设: 血清是有效的.

解 p_1 和 p_2 分别表示 (1) 用血清, (2) 没用血清这两个总体的治愈比例. 我们必须在下面两个假设中作出选择:

$H_0: p_1 = p_2$, 观测到的差异只是出于偶然 (即该血清是无效的).

$H_1: p_1 > p_2$, 该血清是有效的.

在假设 H_0 下

$$\mu_{P_1 - P_2} = 0$$

$$\begin{aligned}\sigma_{P_1 - P_2} &= \sqrt{pq \left(\frac{1}{N_1} + \frac{1}{N_2} \right)} \\ &= \sqrt{0.70 \times 0.30 \times \left(\frac{1}{100} + \frac{1}{100} \right)} = 0.0648\end{aligned}$$

其中我们用两个样本的平均治愈比例 $(75 + 65)/200 = 0.70$ 来估计 p , $q = 1 - p = 0.30$. 则

$$z = \frac{P_1 - P_2}{\sigma_{P_1 - P_2}} = \frac{0.750 - 0.650}{0.0648} = 1.54$$

(a) 在 0.01 的显著性水平下用单边检验, 只有当 z 分数大于 2.33 时才拒绝 H_0 . 因为 z 分数仅有 1.54, 所以我们断定在这个显著性水平下, 结果只是出于偶然.

(b) 在 0.05 的显著性水平下用单边检验, 只有当 z 分数大于 1.645 时才拒绝 H_0 . 所以我们断定在这个显著性水平下, 结果也是出于偶然.

(c) 如果在 0.10 才显著性水平下用单边检验, 只有当 z 分数大于 1.28 时才拒绝 H_0 . 此条件得到了满足, 因此我们断定在 0.10 的水平下, 血清是有效的.

注意, 这些结论都依赖于我们愿意冒多大的风险犯错误. 如果结果只是出于偶然, 而我们断定血清是有效的 (第一类错误), 我们可能会给大量的病人使用这种血清而最终会发现血清是无效的, 这并不是我们愿意承担的风险.

另一方面, 当血清确实有效时我们可能会断定血清是无效的 (第二类错误). 这个结论是非常危险的, 尤其当病人处于生死攸关的时候.

10.21 如果每组有 300 个病人, A 组中有 225 人治愈, B 组中有 195 人治愈. 求解习题 10.20.

解 此时, 两组中病人治愈的比例分别是 $225/300 = 0.750$ 和 $195/300 = 0.650$, 和习题 10.20 中一样. 在假设 H_0 下

$$\mu_{P_1 - P_2} = 0$$

$$\sigma_{P_1 - P_2} = \sqrt{pq \left(\frac{1}{N_1} + \frac{1}{N_2} \right)} = \sqrt{0.70 \times 0.30 \times \left(\frac{1}{300} + \frac{1}{300} \right)} = 0.0374$$

其中, 我们用 $(225 + 195)/600 = 0.70$ 估计 p . 则

$$z = \frac{P_1 - P_2}{\sigma_{P_1 - P_2}} = \frac{0.750 - 0.650}{0.0374} = 2.67$$

因为求出的 z 值大于 2.33, 我们可以在 0.01 的显著性水平下拒绝假设 H_0 ; 也就是说, 我们可以断定该血清是有效的, 而判断错误的概率只有 0.01.

结果表明, 增加样本的容量可以提高决策的可信度. 但是, 在许多情况下增加样本的容量并不可行. 此时, 我们就不得不根据现有的信息作出决策, 而作出错误决策的风险要更大.

10.22 从 A 区和 B 区中分别选取 300 人和 200 人作调查, 发现支持某位候选人的比例分别为 56% 和 48%. 在 0.05 的显著性水平下检验假设: (a) 两个区之间有差异, (b) 此候选人在 A 区更受拥护.

解 p_1 和 p_2 分别表示 A 区和 B 区所有选举人中支持该候选人的比例. 在假设 $H_0: p_1 = p_2$ 下, 我们有

$$\mu_{P_1 - P_2} = 0$$

$$\sigma_{P_1-P_2} = \sqrt{pq \left(\frac{1}{N_1} + \frac{1}{N_2} \right)} = \sqrt{0.528 \times 0.472 \times \left(\frac{1}{300} + \frac{1}{200} \right)} = 0.0456$$

其中, 我们分别用 $(0.56 \times 300 + 0.48 \times 200) / 500 = 0.528$ 和 $1 - 0.528 = 0.472$ 估计 p 和 q . 则

$$z = \frac{P_1 - P_2}{\sigma_{P_1-P_2}} = \frac{0.560 - 0.480}{0.0456} = 1.75$$

(a) 如果我们要判定两个区之间是否有差异, 我们只要在 $H_0: p_1 = p_2$ 和 $H_1: p_1 \neq p_2$ 中作出选择, 这是一个双边检验. 用显著性水平为 0.05 的双边检验, 如果 z 在 -1.96 到 1.96 之外, 我们就拒绝 H_0 . 因为 z 在区间内, 所以在此水平下我们不能拒绝 H_0 , 也就是说, 两个区之间没有显著差异.

(b) 如果我们必须判定此候选人在 A 区是否更受拥护, 我们就要在 $H_0: p_1 = p_2$ 和 $H_1: p_1 > p_2$ 中作出选择, 这是一个单边检验. 用显著性水平为 0.05 的单边检验, 如果 z 大于 1.645, 我们就拒绝 H_0 . 因此, 我们在此水平上拒绝假设 H_0 , 该候选人在 A 区更受拥护.

关于二项分布的检验

10.23 老师给出了 10 道是非题考学生. 为了检验学生是否在猜答案, 老师采用了下面的决策法则:

如果答对了 7 道题或 7 道题以上, 学生就不是全凭猜测.

如果答对的题目少于七道题, 学生就是在猜答案.

求该决策法则犯第一类错误的概率.

解 p 表示答对一题的概率. 10 题中答对 X 题的概率是 $\binom{10}{X} p^X q^{10-X}$, 其中 $q = 1 - p$. 则在 $p = 0.5$ 的假设下 (即学生全凭猜测), 有

$$\begin{aligned} P(\text{答对 7 道题或 7 道题以上}) &= P(\text{答对 7 道题}) + P(\text{答对 8 道题}) \\ &\quad + P(\text{答对 9 道题}) + P(\text{答对 10 道题}) \\ &= \binom{10}{7} \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^3 + \binom{10}{8} \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^2 \\ &\quad + \binom{10}{9} \left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right)^1 + \binom{10}{10} \left(\frac{1}{2}\right)^{10} \\ &= 0.1719 \end{aligned}$$

所以当学生确实在猜答案时, 断定他没有猜的概率是 0.1719. 这就是犯第一类错误的概率.

10.24 在习题 10.23 中, 求当 $p = 0.7$ 时接受假设 $p = 0.5$ 的概率.

解 在假设: $p = 0.7$ 下

$$\begin{aligned} P(\text{答对少于 7 道题}) &= 1 - P(\text{答对 7 道题或 7 道题以上}) \\ &= 1 - \left[\binom{10}{7} \cdot 0.7^7 \cdot 0.3^3 + \binom{10}{8} \cdot 0.7^8 \cdot 0.3^2 \right. \\ &\quad \left. + \binom{10}{9} \cdot 0.7^9 \cdot 0.3 + \binom{10}{10} \cdot 0.7^{10} \right] \\ &= 0.3504 \end{aligned}$$

10.25 在习题 10.23 中, 求当 (a) $p = 0.6$, (b) $p = 0.8$, (c) $p = 0.9$, (d) $p = 0.4$, (e) $p = 0.3$, (f) $p = 0.2$, (g) $p = 0.1$ 时接受假设 $p = 0.5$ 的概率.

表 10.4

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
β	1.000	0.999	0.989	0.945	0.828	0.618	0.350	0.121	0.013

解 (a) 如果 $p = 0.6$, 有

$$\begin{aligned} \text{要求的概率} &= 1 - [P(\text{答对 7 道题}) + P(\text{答对 8 道题}) \\ &\quad + P(\text{答对 9 道题}) + P(\text{答对 10 道题})] \end{aligned}$$

$$= 1 - \left[\binom{10}{7} \cdot 0.6^7 \cdot 0.4^3 + \binom{10}{8} \cdot 0.6^8 \cdot 0.4^2 + \binom{10}{9} \cdot 0.6^9 \cdot 0.4 + \binom{10}{10} \cdot 0.6^{10} \right] = 0.618$$

(b)到(g)的结果可类似得到,相应的值列在了表 10.4 中.注意,在表 10.4 中,求得的概率用 β 表示(犯第二类错误的概率),则 $p = 0.5$ 时 $\beta = 1 - 0.1719 = 0.8281$ (见习题 10.23), $p = 0.7$ 时的 β 值见习题 10.24.

10.26 作习题 10.25 中 β 对 p 的图形,并由此作出习题 10.23 中决策法则的 OC 曲线.

解 要作的图形如图 10-12.注意,和习题 10.14 中的 OC 曲线类似,如果我们作出了 $1 - \beta$ 对 p 的图形,就得到了决策法则的功效曲线.图形说明当真正的 $p \leq 0.4$ 或 $p \geq 0.8$ 时,给定的决策法则对拒绝假设 $p = 0.5$ 是有效的.

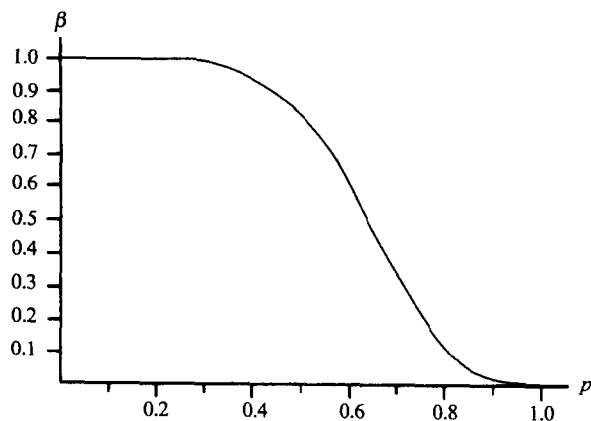


图 10-12

10.27 抛掷一枚硬币 6 次,每次都出现正面.我们能否分别用单边检验和双边检验在(a) 0.05, (b) 0.01 的显著性水平下断定硬币不是均匀的?

解 p 表示抛掷硬币 1 次正面出现的概率.在 $H_0: p = 0.5$ (即硬币是均匀的)的假设下

$$p(X) = P(\text{抛掷硬币 6 次有 } X \text{ 次出现正面}) = \binom{6}{X} \left(\frac{1}{2}\right)^X \left(\frac{1}{2}\right)^{6-X} = \binom{6}{X} \cdot \frac{1}{64}$$

则 1 次, 2 次, 3 次, 4 次, 5 次和 6 次出现正面的概率分别是 $\frac{1}{64}$, $\frac{6}{64}$, $\frac{15}{64}$, $\frac{20}{64}$, $\frac{15}{64}$, $\frac{6}{64}$ 和 $\frac{1}{64}$, 其概率分布图为图 10-13.

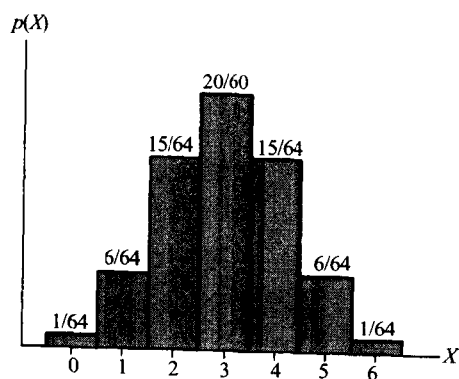


图 10-13

单边检验:

我们要在 $H_0: p = 0.5$ 和 $H_1: p > 0.5$ 中作出选择. 因为 $P(6 \text{ 次出现正面}) = \frac{1}{64} = 0.01562$, $P(5 \text{ 次或 } 6 \text{ 次出现正面}) = \frac{6}{64} + \frac{1}{64} = 0.1094$, 所以我们可以 0.05 的水平下拒绝 H_0 , 但不可以在 0.01 的水平上拒绝 H_0 (即观测到的结果在 0.05 的水平下是显著的, 但在 0.01 的水平下不显著).

双边检验:

我们要在 $H_0: p = 0.5$ 和 $H_1: p \neq 0.5$ 中作出选择. 因为 $P(0 \text{ 次或 } 6 \text{ 次出现正面}) = \frac{1}{64} + \frac{1}{64} = 0.03125$, 所以我们可以 0.05 的水平下拒绝 H_0 , 但不可以在 0.01 的水平下拒绝 H_0 .

10.28 如果有 5 次出现正面, 求解习题 10.27.

解 单边检验: 因为

$$P(5 \text{ 次或 } 6 \text{ 次出现正面}) = \frac{6}{64} + \frac{1}{64} = \frac{7}{64} = 0.1094,$$

所以我们不能在 0.05 或 0.01 的水平下拒绝 H_0 .

双边检验: 因为

$$P(0 \text{ 次或 } 1 \text{ 次或 } 5 \text{ 次或 } 6 \text{ 次出现正面}) = 2 \times \frac{7}{64} = 0.2188,$$

所以我们不能在 0.05 或 0.01 的水平下拒绝 H_0 .

补充习题

用正态分布检验均值和比例

- 10.29** 一个罐子里装有红球和蓝球. 为了检验假设: 红球和蓝球的比例相等, 我们从罐子中有放回地抽取 64 个球, 并记下它们的颜色. 采用如下的决策法则:
如果抽到了 28 至 36 个红球则接受假设. 否则拒绝假设.
- 求拒绝了正确假设的概率.
 - 画图表示决策法则和 (a) 中得到的结果.
- 10.30** (a) 在习题 10.29 中, 如要使拒绝了正确假设的概率不大于 0.01 (即显著性水平为 0.01), 应采用何种决策法则?
(b) 接受假设的显著性水平是多少?
(c) 如果采用 0.05 的显著性水平, 应有何种决策法则?
- 10.31** 假设在习题 10.29 中要检验红球比例大于蓝球这一假设.
- 零假设和备择假设各是什么?
 - 采用单边检验还是双边检验? 为什么?
 - 如果采用 0.05 的显著性水平, 应有何种决策法则?
 - 如果采用 0.01 的显著性水平, 应有何种决策法则?
- 10.32** 抛掷一对骰子 100 次, 点数和为 7 出现了 23 次. 当显著性水平为 0.05 时, 分别用 (a) 双边检验, (b) 单边检验检验骰子是均匀的 (即没有负重) 这一假设. 你认为采用哪一种检验更好? 请说明理由.
- 10.33** 如显著性水平为 0.01, 求解习题 10.32.
- 10.34** 某制造商声称他向一家工厂提供的设备中至少有 95% 符合规格. 从设备中提取一个包含 200 个零件的样本, 发现其中有 18 个次品. 在 (a) 0.01, (b) 0.05 的显著性水平下判断他的声明是否可信.
- 10.35** 一段时间以来, 某所大学物理成绩为 A 的百分比为 10%. 某一学期 300 名学生中有 40 人得 A, 在 (a) 0.05, (b) 0.01 的水平下检验这一结果的显著性.
- 10.36** 某一品牌丝线的平均拉断力为 9.72 盎司, 标准差为 1.40 盎司. 某个含有 36 根线的样品本的平均拉断力为 8.93 盎司. 能否在 (a) 0.05, (b) 0.01 的显著性水平下断定丝线是次品?
- 10.37** 不同学校的许多学生接受了一项测验, 平均分为 74.5 分, 标准差为 8.0 分. 其中一所学校有 200 名学生参加了考试, 平均分为 75.9 分, 在 0.05 的水平下分别用 (a) 单边检验, (b) 双边检验讨论这一结果的显著性, 并说明由此检验得到的结论.
- 10.38** 如果显著性水平是 0.01, 求解习题 10.37.

OC 曲线

- 10.39 参见习题 10.29. 当红球的真正比例是(a)0.6, (b) 0.7, (c) 0.8, (d) 0.9, (e)0.3 时, 求接受红、蓝球比例相等这一假设的概率.
- 10.40 作习题 10.39 的(a) β 对 p , (b) $1 - \beta$ 对 p 的图形. 分别把红球和蓝球类比于硬币的正面和反面, 将此图和习题 10.12 的图相比较.
- 10.41 (a) 如果检验 400 条绳子, 求解习题 10.13 和 10.14.
(b) 当样本容量变大时, 关于犯第二类错误的风险, 你能得出什么结论?
- 10.42 根据习题 10.31 作出(a) OC 曲线, (b) 功效曲线. 并和习题 10.14 中的曲线相比较.

质量控制图

- 10.43 某工厂生产的一种品牌的丝线的平均拉断力为 8.64 盎司, 标准差为 1.28 盎司. 为了判断产品是否符合标准, 每隔 3 小时抽取 16 根丝线, 记下它们的平均拉断力. 在质量控制图上记录(a) 99.73% (或 3σ), (b) 99%, (c) 95% 的控制限, 并解释其应用.
- 10.44 某公司生产的螺栓中大约有 3% 是次品. 为了保证产品的质量, 每隔 4 小时抽取一个含有 200 个螺栓的样本. 求每个样本次品数的(a) 99%, (b) 95% 的控制限. 注意本题只需用到控制上限.

有关均值之差和比例之差的检验

- 10.45 A 工厂生产的 100 个灯泡的平均寿命为 1190 小时, 标准差为 90 小时. B 工厂生产的 75 个灯泡的平均寿命是 1230 小时, 标准差是 120 小时. 在 (a) 0.05, (b) 0.01 的显著性水平下判断两种灯泡的平均寿命是否有差异?
- 10.46 在习题 10.45 中, 在(a) 0.05, (b) 0.01 的显著性水平下检验 B 工厂生产的灯泡优于 A 工厂生产的灯泡这一假设. 说明这题与上题结果的差异. 这题结果是否与上题结果相矛盾?
- 10.47 一所中学进行拼写测验, 32 个男生的平均分是 72 分, 标准差是 8 分. 36 个女生的平均分是 75 分, 标准差是 6 分. 在(a) 0.05, (b) 0.01 的显著性水平下检验假设: 女生的拼写能力强于男生.
- 10.48 为了检验一种新农药对小麦的产量是否有影响, 将一块地分成面积相等的 60 块, 每块地的土壤质量及光照条件等均相同. 对其中 30 块地用新农药, 另 30 块地用旧农药. 用新农药的地上的小麦平均产量是 18.2 蒲式耳, 标准差是 0.63 蒲式耳. 用旧农药的地上的小麦平均产量是 17.8 蒲式耳, 标准差是 0.54 蒲式耳. 在(a) 0.05, (b) 0.01 的显著性水平下检验假设新农药优于旧农药.
- 10.49 A 机器生产的 200 个螺栓样品中有 19 个次品, B 机器生产的 100 个螺栓样品有 5 个次品. 在 0.05 的显著性水平下检验(a) 两台机器所生产的产品质量有差异, (b) B 机器生产的产品质量优于 A 机器的.
- 10.50 A 和 B 两个罐子里装有等量的球, 但是每个罐子里红球与白球的比例未知. 从每个罐子里有放回地抽取 50 个球, 发现从 A 罐中抽到 32 个红球, 从 B 罐中抽到 23 个红球. 在 0.05 的显著性水平下检验 (a) 两个罐子里红球的比例相同, (b) A 罐里红球的比例高于 B 罐.

关于二项分布的检验

- 10.51 参见习题 10.23. 问学生至少应答对多少题才能使老师在(a) 0.05, (b) 0.01, (c) 0.001, (d) 0.06 的显著性水平下求出学生确信学生不是在猜题. 并讨论得到的结果.
- 10.52 类似于习题 10.10, 为习题 10.24 作图.
- 10.53 如果将习题 10.23 决策中的 7 换作 8, 求解习题 10.23 到习题 10.25.
- 10.54 抛掷一枚硬币 8 次有 7 次出现正面. 我们能否在(a) 0.05, (b) 0.10, (c) 0.01 的显著性水平下拒绝硬币是均匀的这一假设? 用双边检验来判定.
- 10.55 用单边检验解习题 10.54.
- 10.56 如果硬币的正面出现 8 次, 求解习题 10.54.
- 10.57 如果硬币的正面出现 6 次, 求解习题 10.54.
- 10.58 一个罐子中有大量的红球和白球. 从中随机抽取 8 个球, 其中有 6 个白球和 2 个红球. 请用适当的检验方法和显著性水平讨论罐子中红球和白球的比例.
- 10.59 讨论如何用抽样理论来调查湖泊里不同种类鱼的比例.

第十一章 小样本理论

小样本

在前面的章节中,我们经常用到这样一个事实,即当样本容量 $N > 30$ 时,我们称之为**大样本**,其中统计量的抽样分布近似为正态分布,且随着 N 的增大,越来越接近正态分布.当样本容量 $N < 30$ 时,我们称之为**小样本**,此时的抽样分布不再能用正态分布来近似,且随着 N 的减少,与正态分布的差别就越来越大,因此对小样本进行适当的修正是必要的.

对小样本统计量的抽样分布的研究称之为**小样本理论**.然而,因所得结论不仅适合小样本问题,而且也适合大样本问题,故有时称之为**精确抽样理论**会更贴切一些.在本章中,我们讨论了三类重要的分布:学生 t 分布, χ^2 分布和 F 分布.

t 分布

我们定义统计量

$$t = \frac{\bar{X} - \mu}{s} \sqrt{N-1} = \frac{\bar{X} - \mu}{s/\sqrt{N}} \quad (1)$$

它是模仿统计量 z 给出的,其中

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

如果我们考虑一个样本容量为 N 的取自均值为 μ 的正态总体(或渐近正态总体)的样本,对每一样本,利用其样本均值 \bar{X} 和样本标准差 s (或 \hat{s})按式(1)来计算 t 值,则对 t 的抽样分布就可以获得(见图 11-1).这个分布可以由下式给出:

$$Y = \frac{Y_0}{\left(1 + \frac{t^2}{N-1}\right)^{N/2}} = \frac{Y_0}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}} \quad (2)$$

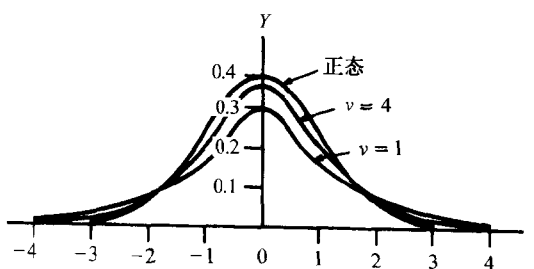


图 11-1 各种 ν 值下的 t 分布

其中 Y_0 是依赖于 N 的常量,使得曲线以下的总面积等于 1, 常量 $\nu = N - 1$ 称之为**自由度**(ν 是希腊字母).

分布(2)被其发现者 W. S. Gossett 称之为**学生分布**或 t 分布. 因他曾在 20 世纪早期以笔名“学生”发表作品而得名.

当 ν 或 N 很大时(当然 $N \geq 30$), 曲线(2)非常近似于标准正态曲线

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

如图 11-1 中所示.

置信区间

正如我们在第九章中所讨论的正态分布一样,利用附录Ⅲ中 t 分布表,我们可以确定 t 分布的 95%, 99% 或其他置信水平的置信区间,由此可以在指定的置信限中估计总体均值 μ .

例如,设 $-t_{0.975}$ 和 $t_{0.975}$ 所对应的 t 值,使得 $(-\infty, -t_{0.975})$ 和 $(t_{0.975}, +\infty)$ 在 t 分布图中各占曲线下总面积的 2.5%, 那么 t 的 95% 置信区间是

$$-t_{0.975} < \frac{\bar{X} - \mu}{s} \sqrt{N-1} < t_{0.975} \quad (3)$$

从(3)式可估计得总体均值 μ 的 95% 置信区间为

$$\bar{X} - t_{0.975} \frac{s}{\sqrt{N-1}} < \mu < \bar{X} + t_{0.975} \frac{s}{\sqrt{N-1}} \quad (4)$$

注意, $t_{0.975}$ 代表 t 分布的 97.5 百分位数, $t_{0.025} = -t_{0.975}$ 代表 2.5 百分位数.

一般来讲,我们可以用式

$$\bar{X} \pm t_c \frac{s}{\sqrt{N-1}} \quad (5)$$

来表示总体均值的置信限,其中的 $\pm t_c$ 称之为**临界值**.它们依赖于置信水平及样本容量,可以在附录Ⅲ中查到.

我们用第九章置信限 $(\bar{X} \pm z_c \sigma / \sqrt{N})$ 和(5)式作比较,可以发现对小样本情形,我们用 t_c 代替 z_c , 用 $\sqrt{N/(N-1)}s = s$ 代替 σ , 其中 s 是 σ 的样本估计值.随着 N 的增加,两种估计都趋于一致.

假设检验和显著性检验

我们在第十章中所讨论的假设检验和显著性检验,或决策规则将很容易推广到小样本的问题中,其惟一的区别就是用 t 值或 t 统计量代替 z 值或 z 统计量.

1. 均值

为检验假设 H_0 : 样本来自均值为 μ 的正态总体,我们用 t 值(或 t 统计量)

$$t = \frac{\bar{X} - \mu}{s} \sqrt{N-1} = \frac{\bar{X} - \mu}{s} \sqrt{N} \quad (6)$$

来进行检验,其中 \bar{X} 是样本容量为 N 的样本均值.这类似于在大样本问题中用 z 值

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}}$$

来进行检验,只是在(6)式中用 $s = \sqrt{N/(N-1)}s$ 代替了 z 值中的 σ .其区别在于 z 服从正态分布,而 t 服从学生分布.随着 N 的增加,它们将趋于一致.

2. 均值之差

假设样本容量为 N_1 和 N_2 的两个随机样本分别取自标准差相等(即 $\sigma_1 = \sigma_2$)的两个正态总体,再假设这两样本的均值和标准差分别为 \bar{X}_1 和 \bar{X}_2 及 s_1 和 s_2 .我们用 t 值来检验假设 H_0 : 样本取自同一总体(即 $\mu_1 = \mu_2$, $\sigma_1 = \sigma_2$), 其中

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma \sqrt{1/N_1 + 1/N_2}} \quad \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}} \quad (7)$$

t 服从自由度为 $\nu = N_1 + N_2 - 2$ 的学生分布.(7)式类似于第十章(2)式中的 z 值,只是其中 $\sigma_1 = \sigma_2 = \sigma$, 用加权平均值

$$\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{(N_1 - 1) + (N_2 - 1)} = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}$$

来估计 σ^2 , 其中 s_1^2 和 s_2^2 分别是 σ_1^2 和 σ_2^2 的无偏估计.

χ^2 分布

设统计量

$$\chi^2 = \frac{Ns^2}{\sigma^2} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_N - \bar{X})^2}{\sigma^2} \quad (8)$$

其中 χ 是希腊字母, 读作“卡”, χ^2 读作“卡方”.

如果我们考虑取自标准差为 σ 的正态总体的容量为 N 的样本, 对每一样本, 我们计算其 χ^2 值, 则 χ^2 的抽样分布即可获得. 这个分布我们称之为 χ^2 分布. 该分布由下式给出:

$$Y = Y_0(\chi^2)^{\frac{1}{2}(\nu-2)} e^{-\frac{1}{2}\chi^2} = Y_0 \chi^{\nu-2} e^{-\frac{1}{2}\chi^2} \quad (9)$$

其中 $\nu = N - 1$ 是自由度, Y_0 是依赖于 ν 的常量, 使得曲线下总面积为 1. 对应不同的 ν 值有不同的 χ^2 曲线图 (见图 11-2), 其中 Y 在 $\chi^2 = \nu - 2 (\nu \geq 2)$ 时取最大值.

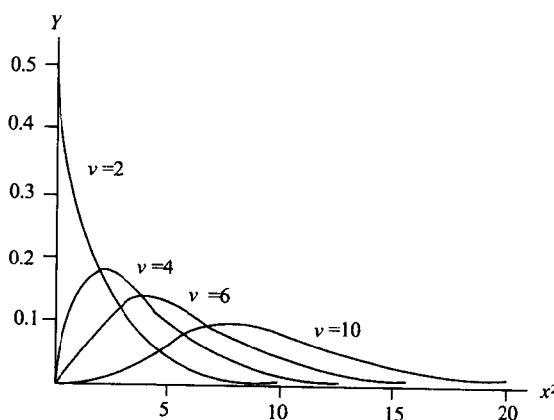


图 11-2 各种 ν 值的 χ^2 分布

χ^2 的置信区间

正如前面对正态分布和 t 分布所讨论的一样, 我们可以利用附录 IV 中的 χ^2 分布表确定 χ^2 的 95%, 99% 或其他置信水平的置信限或置信区间. 下面我们根据样本标准差 s , 在一特定置信限内估计总体标准差 σ .

例如, 设 $\chi_{0.025}^2$ 和 $\chi_{0.975}^2$ 是 χ^2 的使得区间 $(0, \chi_{0.025}^2)$ 和 $(\chi_{0.975}^2, +\infty)$ 所对应的 χ^2 分布曲线下方的面积各占总面积的 2.5% 的值 (称为临界值), 那么 χ^2 的 95% 置信区间为

$$\chi_{0.025}^2 < \frac{Ns^2}{\sigma^2} < \chi_{0.975}^2 \quad (10)$$

由此我们可得 σ 的 95% 估计区间为

$$\frac{s\sqrt{N}}{\chi_{0.975}} < \sigma < \frac{s\sqrt{N}}{\chi_{0.025}} \quad (11)$$

对其他的置信区间我们同样可得, 其中 $\chi_{0.025}$ 和 $\chi_{0.975}$ 分别代表 χ^2 分布的 2.5 和 97.5 百分位数的平方根.

附录 IV 给出了 χ^2 分布的相应于自由度 ν 的百分位数. 当 $\nu \geq 30$ 时, $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ 的分布接近于标准正态分布, 因此当 $\nu \geq 30$ 时, 我们可以直接利用正态分布表来查临界值. 若 χ_p^2 和 z_p 分别表示 χ^2 和正态分布的第 100 p 个百分位数, 我们有

$$\chi_p^2 = \frac{1}{2}(z_p + \sqrt{2\nu - 1})^2 \quad (12)$$

这种情形中所得结论与第八、第九章中所得结论是一致的。

在第 12 章中将对 χ^2 分布的更广泛的应用。

自由度

为了计算(1)式和(8)式的统计量,我们有必要利用某些总体参数和从样本中所得到的观察值.如果这些参数是未知的,他们将必须利用样本来估计。

统计量的**自由度**(一般以 ν 记号)定义为样本中独立观察值的个数(即样本容量) N 减去 K , 其中 K 是要由样本来估计的总体参数个数, 记为 $\nu = N - K$ 。

在统计量(1)式中,样本容量为 N ,由样本可计算 \bar{X} 和 s .但我们要估计 μ 值,被估参数只有 μ 一个,因此 $K=1$,故 $\nu = N - 1$ 。

在统计量(8)式中,样本容量为 N ,由样本可计算 s .但我们要估计 σ 值,被估参数只有 σ 一个,因此 $K=1$,故 $\nu = N - 1$ 。

F 分布

由前几章可见,有时候知道某两样本均值之差 $\bar{X}_1 - \bar{X}_2$ 的抽样分布在某些应用中是非常重要的.同样,我们将需要知道某两样本方差之差 $S_1^2 - S_2^2$ 的抽样分布.然而,事实已证明这个分布的计算比较困难.因此,我们将考虑比值 S_1^2/S_2^2 ,因为比值的大或小都表示了两方差相差之大,而若比值接近 1 则表示了两方差很接近.两样本方差比值的抽样分布是可以求出的,我们称之为 **F 分布**。

为了更精确表示,我们假设有两个容量为 N_1 和 N_2 分别取自方差为 σ_1^2 和 σ_2^2 的正态(或渐近正态)总体的样本,我们定义统计量

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} = \frac{N_1 S_1^2 / [(N_1 - 1)\sigma_1^2]}{N_2 S_2^2 / [(N_2 - 1)\sigma_2^2]} \quad (13)$$

其中

$$\hat{S}_1^2 = \frac{N_1 S_1^2}{N_1 - 1} \quad \hat{S}_2^2 = \frac{N_2 S_2^2}{N_2 - 1} \quad (14)$$

则 F 的抽样分布称之为 Fisher 的 F 分布,或简称为 **F 分布**,其自由度为 $\nu_1 = N_1 - 1$ 和 $\nu_2 = N_2 - 1$.这个分布由下式给出:

$$Y = \frac{CF^{(\nu_1/2)-1}}{(\nu_1 F + \nu_2)^{(\nu_1+\nu_2)/2}} \quad (15)$$

其中 C 是依赖于 ν_1 和 ν_2 的常数,使得曲线下的总面积为 1.其图形类似于图 11-3,但随着 ν_1 和 ν_2 的变化,曲线的形状也发生较大的变化。

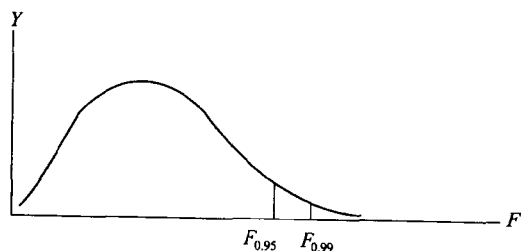


图 11-3

附录 V 和附录 VI 给出了 F 分布的百分位数,其中 $F_{0.95}$ 和 $F_{0.99}$ 分别表示使得 F 分布在 $(F_{0.95}, +\infty)$ 和 $(F_{0.99}, +\infty)$ 内所对应的面积占总面积分别为 5%, 1% 的 F 分布的百分位数.对假设检验问题,它们代表了 5% 和 1% 的显著性水平的临界值.由此可确定,方差 S_1^2 是否显

著大于 S_2^2 . 在实际中, 一般选方差较大的样本为样本 1.

习题及解答

t 分布

- 11.1 自由度为 9 的 t 分布由图 11-4 给出, 求 t_1 值使得 (a) 右边阴影部分面积为 0.05, (b) 总的阴影部分面积为 0.05, (c) 总的非阴影部分面积为 0.99, (d) 左边阴影部分面积为 0.01, (e) t_1 左边的面积为 0.90.

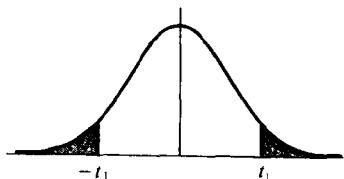


图 11-4

解 (a) 若右边阴影部分面积为 0.05, 则 t_1 左边的面积为 $1 - 0.05 = 0.95$, t_1 代表 t 分布的第 95 个百分位数即 $t_{0.95}$. 由附录 III, 对应自由度 9 和 $t_{0.95}$, 找到 t_1 的值为 1.83.
(b) 若总阴影部分面积为 0.05, 则由对称性可知右边阴影部分面积为 0.025. 因此 t_1 左边的面积为 $1 - 0.025 = 0.975$. 同 (a) 可由附录 III, 找出 t_1 的值为 2.26.
(c) 总的非阴影部分面积为 0.99, 则总阴影部分面积为

0.01, t_1 左边的面积为 $1 - 0.01/2 = 0.995$, 从附录 III 知 $t_1 = 3.25$.

(d) 利用对称性可知右边阴影部分面积为 0.01, 由附录 III, 可知 $t_{0.99} = 2.82$, 故 $t_1 = -2.82$.

(e) 利用附录 III, 知 $t_{0.90} = 1.38$, 即 $t_1 = 1.38$.

- 11.2 求 t 的临界值, 使 t 分布的右尾面积为 0.05, 若自由度 ν 为 (a) 16, (b) 27, (c) 200.

解 利用附录 III, 可找到 $t_{0.95}$ 的值 (a) $\nu = 16$ 时, $t_{0.95} = 1.75$; (b) $\nu = 27$ 时, $t_{0.95} = 1.70$; (c) $\nu = 200$ 时, $t_{0.95} = 1.645$ (最后一值可在正态曲线图找到, 在附录 III 中, 相应于 ν 取无穷大).

- 11.3 正态分布的 95% 的临界值 (双尾) 为 ± 1.96 , 则相应的 t 分布的临界值是多少? 若 (a) $\nu = 9$, (b) $\nu = 20$, (c) $\nu = 30$, (d) $\nu = 60$.

解 对 95% 的临界值 (双尾), 在图 11-4 中对应的阴影部分面积是 0.05, 因此右尾阴影部分面积为 0.025, 相应的 t 的临界值为 $t_{0.975}$, 故所求临界值为 $\pm t_{0.975}$. 对给定的 ν 值, 得相应的 $\pm t_{0.975}$ 值分别为: (a) ± 2.26 , (b) ± 2.09 , (c) ± 2.04 , (d) ± 2.00 .

- 11.4 对某个球面的直径, 给出了含有 10 个测量值的一个样本, 其均值 $\bar{X} = 438$ 厘米, 标准差 $s = 0.06$ 厘米, 求实际直径的 (a) 95%, (b) 99% 的置信限.

解 (a) 95% 的置信限为 $\bar{X} \pm t_{0.975}(s/\sqrt{N-1})$.

因为 $\nu = N - 1 = 10 - 1 = 9$, 可求得 $t_{0.975} = 2.26$ (见习题 11.3(a)). 再将 $\bar{X} = 4.38$, $s = 0.06$ 代入, 即求得 95% 置信限为 $4.38 \pm 2.26(0.06/\sqrt{10-1}) = 4.38 \pm 0.0452$, 即我们有 95% 的把握认为直径的均值处于 4.335 厘米与 4.425 厘米之间.

(b) 99% 置信限为 $\bar{X} \pm t_{0.995}(s/\sqrt{N-1})$. 对 $\nu = 9$, $t_{0.995} = 3.25$, 则 99% 的置信限为

$$4.38 \pm 3.25(0.06/\sqrt{10-1}) = 4.38 \pm 0.0650,$$

因此 99% 的置信区间为 (4.315 厘米, 4.445 厘米).

- 11.5 对 25 个随机选取的工人记录其因与工作有关的腕骨综合症而离岗的天数, 结果由表 11.1 给出. 当这些数据用来对所有与工作有关的腕骨综合症的总体的期望值设立一个置信区间时, 必须有一基本假设: 即这些总体的离岗天数是服从正态分布的. 用这些数据来检验正态假设. 若假设成立, 则设立 μ 的 95% 的置信区间.

解 由 Minitab 软件产生的正态概率图 (见图 11-5) 表明正态性假设是合理的, 此因 p -值大于 0.15. p -值可用来检验零假设: 数据取自正态分布总体. 如果显著性水平为 0.05, 那么当 p -值小于 0.05 时才拒绝总体分布是正态这个假设. 由于用 Kolmogorov-Smirnov 检验得到的 p -值大于 0.05, 因此我们不能拒绝正态性假设.

表 11.1

21	23	33	32	37
40	37	29	23	29
24	32	24	46	32
17	29	26	46	27
36	38	28	33	18

用 Minitab 软件可得到置信区间. 总体均值的 95% 置信区间为每年 27.21 至 33.59 天.

MTB> interval 95 % confidence for data in c1

Confidence Intervals

Variable	N	Mean	StDev	SE Mean	95.0 % CI
days	25	30.40	7.72	1.54	(27.21, 33.59)

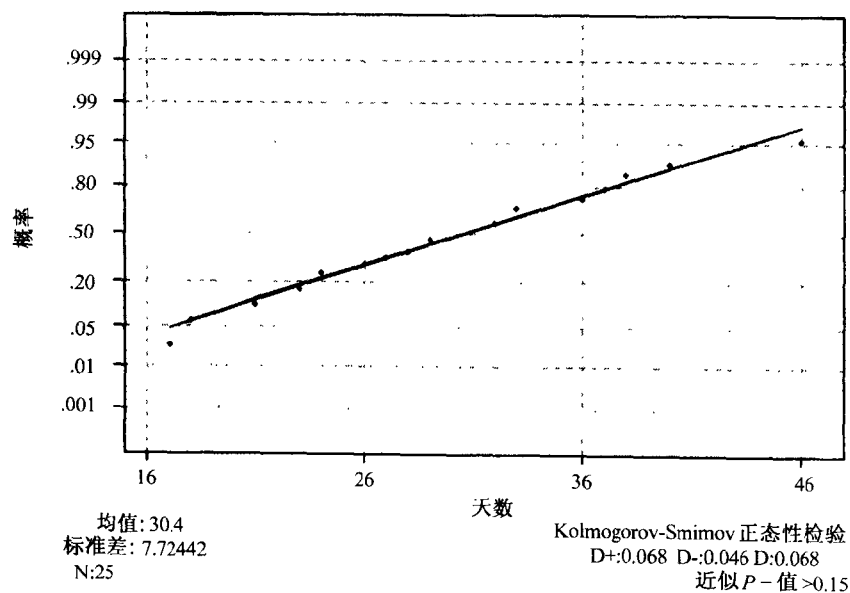


图 11-5 正态概率图

- 11.6 一台机器生产厚度为 0.050 英寸的垫圈. 为确定该机器是否仍正常运转, 选了 10 个垫圈, 测得其厚度均值为 0.053 英寸, 标准差为 0.003 英寸. 在显著性水平 (a) 0.05 和 (b) 0.01 下来检验假设: 机器仍正常运转.

解 我们将在下列假设中做出选择:

$H_0: \mu = 0.050$, 即机器正常运转.

$H_1: \mu \neq 0.050$, 即机器不能正常运转.

因此, 需用到双边检验. 在假设 H_0 下, 我们有

$$t = \frac{\bar{X} - \mu}{s} \sqrt{N - 1} = \frac{0.053 - 0.050}{0.003} \sqrt{10 - 1} = 3.00$$

(a) 对于显著性水平为 0.05 的双边检验, 我们采用决策规则:

若 t 值处于区间 $(-t_{0.975}, t_{0.975})$ 之间, 则接受 H_0 , 否则拒绝 H_0 . 对自由度为 $10 - 1 = 9$ 而言, 该区间为 $(-2.26, 2.26)$. 因为 $t = 3.00$ 在该区间之外, 故在 0.05 水平下拒绝 H_0 .

(b) 对于显著性水平为 0.01 的双边检验, 我们采用决策规则:

若 t 值处于区间 $(-t_{0.995}, t_{0.995})$ 之间, 则接受 H_0 , 否则拒绝 H_0 . 自由度为 9 时, 该区间为

$(-3.25, 3.25)$. 因为 $t = 3.00$ 在该区间之内, 故在 0.01 水平下, 我们接受 H_0 .

因为在水平 0.05 时拒绝 H_0 , 而在水平 0.01 时接受 H_0 , 故我们说该样本结论是**可能显著**(见习题 10.5 最后的专用名词). 因此有必要检查一下该机器或者至少再抽样检查一次.

- 11.7** 一个经营铁锤的经理进行了一项检验: 零假设 $\mu = 50$ 美元相对于备择假设 $\mu \neq 50$ 美元, 其中 μ 表示顾客购买铁锤时所付的平均价格. 对 28 个顾客的统计数据如表 11.2 所示. 用 t 分布来进行假设检验, 设这些数据选自一服从正态分布的总体. 此正态假设可由各种不同的**正态性检验**检测出来. Minitab 给出了三种不同的正态检验. 在惯用的显著性水平 $\alpha = 0.05$ 下进行正态性检验, 若正态假设成立, 我们再继续显著性水平 $\alpha = 0.05$ 下检验假设 $H_0: \mu = 50$ 美元对 $H_1: \mu \neq 50$ 美元.

表 11.2

68	49	45	76	65	50
54	92	24	36	60	66
57	74	52	75	36	40
62	56	94	57	64	
72	65	59	45	33	

解 由 Minitab 软件给出的 Anderson-Darling 正态性检验的 p -值 = 0.922, Ryan-Joyner 正态性检验的 p -值大于 0.1, Kolmogorov-Smirnov 正态性检验的 p -值大于 0.15. 在这三种情况下, 对显著性水平 5%, 都不拒绝数据来自正态总体这个假设, 这是因为零假设当且仅当显著性水平大于 p -值时被拒绝. 对每个顾客所付的平均价格的检验由 Minitab 软件给在下面. 如用经典的检验假设的方法, 那么当检验统计量的绝对值大于 2.05 时拒绝零假设. 临界值 2.05 由自由度为 27 的 t 分布查得. 因经计算得检验统计量的值为 18.5, 故我们拒绝零假设, 而认为所付的平均价格不是 50 美元. 如果用 p -值来检验假设, 则因 p -值 = 0.0000 小于显著性水平 0.05, 我们也拒绝零假设.

Data Display

Amount

```
68 54 57 62 72 49 92 74 56
65 45 24 52 94 59 76 36 75
57 45 65 60 36 64 33 50 66
40
```

```
MTB> TTest 0, 0, 'Amount';
```

```
SUBC> Alternative 0.
```

T-Test of the Mean

Test of $\mu = 0.00$ vs $\mu \neq 0.00$

Variable	N	Mean	StDev	SE Mean	T	P
Amount	28	58.07	16.61	3.14	18.50	0.0000

- 11.8** 从一个城市的某个地区所选的 16 名学生的 IQ 均值为 107, 标准差为 10, 而该城市另一地区所选的 14 名学生的 IQ 均值为 112, 标准差为 8, 则问在显著性水平为 (a) 0.01 和 (b) 0.05 下, 这两组学生的 IQ 值是否存在显著差异?

解 若令 μ_1 和 μ_2 分别记为两个地区学生的总体 IQ 均值, 则我们将从下面假设中做出选择:

$H_0: \mu_1 = \mu_2$, 即两组学生的 IQ 没差异.

$H_1: \mu_1 \neq \mu_2$, 即存在差异.

在假设 H_0 下,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma \sqrt{1/N_1 + 1/N_2}}, \quad \text{其中} \quad \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}}$$

因此

$$\sigma = \sqrt{\frac{16 \times 10^2 + 14 \times 8^2}{16 + 14 - 2}} = 9.44, \quad t = \frac{112 - 107}{9.44 \sqrt{1/16 + 1/14}} = 1.45$$

(a) 在显著性水平 0.01 下用双边检验, 若 t 值处于区间 $(-t_{0.995}, t_{0.995})$ 之间, 则接受 H_0 , 否则拒绝 H_0 . 而对自由度为 $N_1 + N_2 - 2 = 28$ 而言, 该区间为 $(-2.76, 2.76)$, 故接受 H_0 .

(b) 类似于 (a) 可得在水平 0.05 下也接受 H_0 .

由此得出结论, 两组学生的 IQ 值无显著差异.

- 11.9** 对 15 名随机抽取的公立大学学生和 10 名随机抽取的私立大学学生, 统计了他们每年学费、住宿、伙食的开支情况, 记录在表 11.3 中. 检验零假设: 私立学生每年的平均开支比公立学生多 1 万元, 对应的备择假设: 平均开支数相差不是 1 万元. 在进行零假设检验之前, 先在显著性水平 0.05 下检验正态假设和方差相等的假设.

表 11.3

公立大学			私立大学	
4.2	9.1	11.6	13.0	17.7
6.1	7.7	10.4	18.8	17.6
4.9	6.5	5.0	13.2	19.8
8.5	6.2	10.4	14.4	16.8
4.6	10.2	8.1	17.7	16.1

解 对公立大学的数据, 由 Minitab 软件给出的 Anderson-Darling 正态性检验显示在图 11-6 中. 因 p -值(0.432)不小于显著性水平(0.05), 故不拒绝正态性假设.

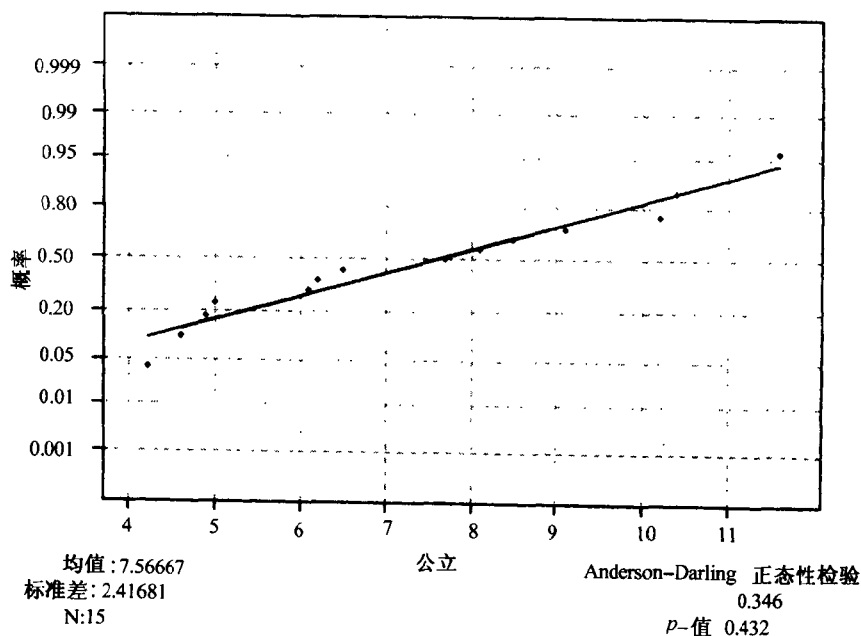


图 11-6 正态概率图

类似地对私立大学的数据, Anderson-Darling 正态性检验的 p -值为 0.394. 因 p -值不小于 0.05, 故对私立大学的数据也不拒绝正态性假设. Bartlett 和 Levene 的方差相等检验的 p -值分别为 0.885 和 0.651, 这表明可认为两总体方差相等. 对这个检验 Minitab 输出如下:

Homogeneity of Variance

Bartlett's Test(normal distribution)

Test Statistic : 0.021
P-Value : 0.885
Levene's Test(any continuous distribution)
Test Statistic : 0.210
P-Value : 0.651

因均值差的 95% 置信区间包含 -10, 因此在显著性水平 0.05 下不拒绝零假设.

Row	Public	Private
1	4.2	13.0
2	6.1	18.8
3	4.9	13.2
4	8.5	14.4
5	4.6	17.7
6	9.1	17.7
7	7.7	17.6
8	6.5	19.8
9	6.2	16.8
10	10.2	16.1
11	11.6	
12	10.4	
13	5.0	
14	10.4	
15	8.1	

MTB>TwoSample95.0'Public''Private';
SUBC>Alternative 0;
SUBC>Pooled.

Two Sample T-Test and Confidence Interval

Two sample T for Public vs Private

	N	Mean	StDev	SE Mean
Public	15	7.57	2.42	0.62
Private	10	16.51	2.31	0.73

95% CI for mu Public-mu Private: (-10.95, -6.94)

χ^2 分布

11.10 自由度为 5 的 χ^2 分布如图 11-7 所示. 求相应的 χ^2 的临界值, 使得 (a) 右边阴影部分面积为 0.05, (b) 总阴影部分面积为 0.05, (c) 左边阴影部分面积为 0.10, (d) 右边阴影部分面积为 0.01.

解 (a) 若右边阴影部分面积为 0.05, 则 χ^2 左边面积为 $1 - 0.05 = 0.95$, χ^2 代表 χ^2 分布的第 95 个百分位数, 即 $\chi^2_{0.95}$. 参照附录 IV, 对应于自由度 5 和 $\chi^2_{0.95}$, 可得 χ^2 临界值为 11.1.

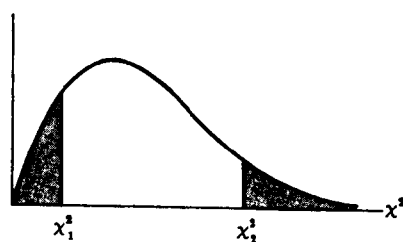


图 11-7

(b) 因为分布是非对称的, 故对总阴影部分面积为 0.05 求临界值有很多解. 例如, 设右边阴影部分面积为 0.04, 则左边阴影面积为 0.01. 但是, 特别说明, 一般指两边阴影面积是相等的. 在这里, 即为左、右阴影面积均为 0.025.

若右边阴影面积为 0.025, 则同 (a), 由附录 IV 可得临界值 $\chi^2_2 = 12.8$. 同样若左边阴影面积为 0.025, 则由附录 IV, 找对应自由度 5 和 $\chi^2_{0.025}$ 的值, 即得临界值 $\chi^2_1 = 0.831$.

(c) $\chi_1^2 = 1.61$.

(d) $\chi_2^2 = 15.1$.

- 11.11 对应自由度 ν 为(a) 15, (b) 21, (c) 50, 求 χ^2 的临界值, 使得 χ^2 分布的右侧尾面积为 0.05.

解 由附录 IV, 对应于 $\chi_{0.95}^2$, 可得临界值为(a) 25.0, $\nu = 15$; (b) 32.7, $\nu = 21$; (c) 67.5, $\nu = 50$.

- 11.12 对应自由度为(a) 9, (b) 28, (c) 40, 求 χ^2 的中位数.

解 由附录 IV, 对应于 $\chi_{0.50}^2$, 可得中位数为(a) 8.34, $\nu = 9$; (b) 27.3, $\nu = 28$; (c) 39.3, $\nu = 40$.

我们注意到一个有趣现象, 即所求中位数很接近相应的自由度. 事实上, 对 $\nu > 10$, 其相应中位数等于 $\nu - 0.7$.

- 11.13 在一个学校 1000 名男生中随机选 16 名男生, 测得他们身高的标准差为 2.40 英寸. 求该校所有男生身高标准差的(a) 95%, (b) 99% 的置信限.

解 (a) 95% 置信限由 $s\sqrt{N}/\chi_{0.975}$ 和 $s\sqrt{N}/\chi_{0.025}$ 给出.

对 $\nu = 16 - 1 = 15$, $\chi_{0.975}^2 = 27.5$ (或者 $\chi_{0.975} = 5.24$), $\chi_{0.025}^2 = 6.26$ (或者 $\chi_{0.025} = 2.50$), 则 95% 置信限为 $2.40 \times \sqrt{16}/5.24$ 和 $2.40 \times \sqrt{16}/2.50$ (即 1.83 和 3.84 英寸), 因此我们说总体标准差的 95% 置信区间为 (1.83, 3.84).

(b) 99% 置信限由 $s\sqrt{N}/\chi_{0.995}$ 和 $s\sqrt{N}/\chi_{0.005}$ 给出. 类似于(a) 可得总体标准差的 99% 置信区间为 (1.68, 4.49).

- 11.14 对自由度(a) $\nu = 50$, (b) $\nu = 100$ 求 $\chi_{0.95}^2$.

解 当 $\nu > 30$ 时, $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ 的分布非常接近标准正态分布. 若令 z_p 表示标准正态分布的下 p 分位数, 则在很大程度上, 近似地有:

$$\sqrt{2\chi_p^2} - \sqrt{2\nu - 1} = z_p \quad \sqrt{2\chi_p^2} = z_p + \sqrt{2\nu - 1}$$

因此

$$\chi_p^2 = \frac{1}{2}(z_p + \sqrt{2\nu - 1})^2$$

(a) 若 $\nu = 50$, 则 $\chi_{0.95}^2 = \frac{1}{2}(z_{0.95} + \sqrt{2 \times 50 - 1})^2 = \frac{1}{2}(1.64 + \sqrt{99})^2 = 67.2$, 这个值与附录 IV 中所查值 67.5 非常接近.

(b) 若 $\nu = 100$, 则 $\chi_{0.95}^2 = \frac{1}{2}(z_{0.95} + \sqrt{2 \times 100 - 1})^2 = \frac{1}{2}(1.64 + \sqrt{199})^2 = 124.0$ (实际值 = 124.3).

- 11.15 200 个电灯泡寿命的样本标准差为 100 小时, 求所有这类电灯泡寿命的标准差的(a) 95%, (b) 99% 的置信限.

解 类似于习题 11.14 可得 $\chi_{0.975}^2 = 239$, $\chi_{0.025}^2 = 161$, $\chi_{0.995}^2 = 253$, $\chi_{0.005}^2 = 150$, 即 $\chi_{0.975} = 15.5$, $\chi_{0.025} = 12.7$, $\chi_{0.995} = 15.9$, $\chi_{0.005} = 12.2$. 再类似于习题 11.13 可得:

(a) 总体标准差 95% 的置信区间为 (91.2, 111.3).

(b) 总体标准差 99% 的置信区间为 (88.9, 115.9).

- 11.16 一轮轴制造工在制造过程中必须保持轮轴的平均直径为 5.000 厘米. 此外, 为了保证轮轴与轮子相配, 直径的标准差必须小于或等于 0.005 厘米. 现抽取 20 个轮轴, 它们的直径由表 11.4 给出.

表 11.4

4.996	4.998	5.002	4.999
5.010	4.997	5.003	4.998
5.006	5.004	5.000	4.993
5.002	4.996	5.005	4.992
5.007	5.003	5.000	5.000

制造工希望检验零假设: 总体标准差为 0.005 厘米, 对应的备择假设: 总体标准差

超过 0.005 厘米. 若备择假设成立, 则必须停止制造过程, 对机器进行修理. 在检验程序中, 我们假设轮轴直径是服从正态分布的. 在显著水平 0.05 下检验这个正态假设, 若该假设成立, 再在 0.05 显著水平下检验关于总体标准差的零假设.

解 类似于 Shapiro-Wilk 正态性检验的 Ryan-Joiner 正态性检验显示在图 11-8, 其 p -值超过 0.10, 因此在显著性水平 0.05 下不拒绝正态性假设.

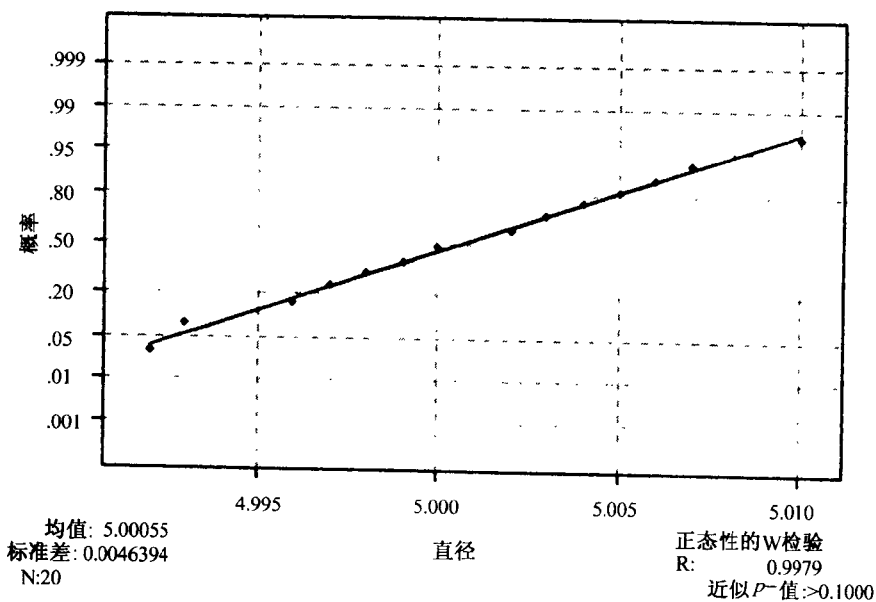


图 11-8 正态概率图

对除数为 $N-1$ 的标准差, Minitab 软件计算结果如下:

Data Display

Diameter

4.996	5.010	5.006	5.002	5.007	4.998	4.997
5.004	4.996	5.003	5.002	5.003	5.000	5.005
5.000	4.999	4.998	4.993	4.992	5.000	

MTB>standard deviation cl

Column Standard Deviation

Standard deviation of Diameter = 0.0046394

总体标准差小于或等于 0.005 的检验即要求我们在下面两个假设中作出选择:

$H_0: \sigma = 0.005$, 即观察结果纯属偶然.

$H_1: \sigma > 0.005$, 即有很大变化.

样本的 χ^2 值为

$$\chi^2 = \frac{(N-1)S^2}{\sigma^2} = \frac{19 \times 0.0046394^2}{0.005^2} = 16.4$$

使用单边检验, 对于自由度为 19 及显著性水平为 0.05, 如果 χ^2 的值大于 $\chi_{0.95}^2 = 30.1$, 则拒绝 H_0 . 因此在显著性水平 0.05 下我们不拒绝 H_0 .

- 11.17** 以往, 一装有某机器的 40 盎司包裹, 其重量的标准差为 0.25 盎司. 现随机抽样 20 个包裹, 其标准差为 0.32 盎司, 问在显著性水平 (a) 0.05, (b) 0.01 下是否有明显增加?

解 我们必须在以下假设中做出选择:

$H_0: \sigma = 0.25$ 盎司, 即观察结果纯属偶然.

$H_1: \sigma > 0.25$ 盎司, 即有明显变化.

其样本的 χ^2 值为

$$\chi^2 = \frac{Ns^2}{\sigma^2} = \frac{20 \times 0.32^2}{0.25^2} = 32.8$$

(a) 用单边检验, 在显著性水平 0.05 下, 若样本 χ^2 值大于 $\chi_{0.95}^2$, 则拒绝 H_0 . 对自由度 $\nu = 20 - 1 = 19$, $\chi_{0.95}^2 = 30.1$. 因此在 0.05 显著性水平下, 我们必须拒绝 H_0 .

(b) 用单边检验, 在显著水平 0.01 下, 若样本 χ^2 值大于 $\chi_{0.99}^2 = 36.2$, 则拒绝 H_0 . 因此, 我们在 0.01 显著性水平下, 支持 H_0 (或者说接受 H_0).

由上, 我们得出结论: 可能有增加, 故应该对机器进行检测.

F 分布

11.18 从两个服从正态且方差为 16 和 25 的总体中分别抽取样本容量为 9 和 12 的样本. 若样本方差分别为 20 和 8, 试确定在显著性水平 (a) 0.05, (b) 0.01 下, 第一个样本的方差是否比第二个样本方差明显偏大?

解 对样本 1 和样本 2, 有 $N_1 = 9$, $N_2 = 12$, $\sigma_1^2 = 16$, $\sigma_2^2 = 25$, $s_1^2 = 20$, $s_2^2 = 8$, 因此

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{N_1 S_1^2 / [(N_1 - 1)\sigma_1^2]}{N_2 S_2^2 / [(N_2 - 1)\sigma_2^2]} = \frac{9 \times 20 / [(9 - 1) \times 16]}{12 \times 8 / [(12 - 1) \times 25]} = 4.03$$

(a) F 的分子、分母的自由度分别为 $\nu_1 = N_1 - 1 = 8$, $\nu_2 = N_2 - 1 = 11$. 从附录 V, 可以找到 $F_{0.95} = 2.95$. 由于 $F = 4.03 > F_{0.95}$, 故得结论: 在显著性水平 0.05 下, 样本 1 的方差明显大于样本 2 的方差.

(b) 对 $\nu_1 = 8$, $\nu_2 = 11$ 由附录 V, 可得 $F_{0.99} = 4.74 > 4.03 = F$, 因此得结论: 在显著性水平 0.01 下, 样本 1 方差不明显大于样本 2 方差.

11.19 从两个服从正态分布且方差分别为 20 和 36 的总体中分别抽取样本容量为 8 和 10 的样本. 求样本 1 方差大于两倍样本 2 方差的概率.

解 已知 $N_1 = 8$, $N_2 = 10$, $\sigma_1^2 = 20$, $\sigma_2^2 = 36$ 故

$$F = \frac{8S_1^2/7 \times 20}{10S_2^2/9 \times 36} = 1.85 \frac{S_1^2}{S_2^2}$$

F 的分子、分母的自由度分别为 $\nu_1 = N_1 - 1 = 7$, $\nu_2 = N_2 - 1 = 9$. 若 S_1^2 大于 $2S_2^2$, 则

$$F = 1.85 \frac{S_1^2}{S_2^2} > 1.85 \times 2 = 3.70$$

从附录 V 和 VI 中可知: $0.01 < P(F > 3.70) < 0.05$. 为得更确切的值, 我们必须有一个 F 分布的更精确的表.

补充习题

t 分布

11.20 对一自由度为 15 的 t 分布, 求 t_1 的值使得 (a) t_1 右边的面积为 0.01, (b) t_1 左边的面积为 0.95, (c) t_1 右边的面积为 0.10, (d) t_1 右边和 $-t_1$ 左边的总面积为 0.01, (e) $-t_1$ 与 t_1 之间的面积为 0.95.

11.21 对自由度 ν 为 (a) 4, (b) 12, (c) 25, (d) 60, (e) 150, 分别求其相应的 t 的临界值, 使得该分布的右尾面积是 0.01.

11.22 求 t 分布的值 t_1 , 使其分别满足下列条件:

(a) $-t_1$ 与 t_1 之间的面积为 0.90, $\nu = 25$.

(b) $-t_1$ 左边的面积为 0.025, $\nu = 20$.

(c) t_1 右边与 $-t_1$ 左边的总面积为 0.01, $\nu = 5$.

(d) t_1 右边的面积是 0.55, $\nu = 16$.

11.23 若一变量 U 服从自由度为 10 的 t 分布, 求常数 C , 使得: (a) $P(U > C) = 0.05$, (b) $P(-C \leq U \leq C) = 0.98$, (c) $P(U \leq C) = 0.20$, (d) $P(U \geq C) = 0.90$.

11.24 正态分布的 99% 临界值 (双边) 为 ± 2.58 , 求相应 t 分布的临界值, 若 (a) $\nu = 4$, (b) $\nu = 12$, (c) $\nu = 25$, (d) $\nu = 30$, (e) $\nu = 40$.

- 11.25 取 12 根纱线样品,测量它们的抗断强度,可得其均值为 7.38 克,标准差为 1.24 克,求实际抗断强度的 (a) 95%, (b) 99% 置信限.
- 11.26 假设大样本理论方法对习题 11.25 适用,比较其所得结果.
- 11.27 对 5 个不同个体测量他们对某种刺激的反应时间分别为:0.28 秒,0.30 秒,0.27 秒,0.33 秒,0.31 秒.求实际反应时间的 (a) 95%, (b) 99% 的置信限.
- 11.28 以往某公司生产的电灯泡的平均寿命为 1120 小时,标准差为 125 小时,现从一批新生产的灯泡中抽取 8 个样品,得其平均寿命为 1070 小时,分别在显著性水平 (a) 0.05, (b) 0.01 下检验假设:灯泡的平均寿命没变.
- 11.29 在问题 11.28 中,在显著性水平 (a) 0.05, (b) 0.01 下,检验零假设: $\nu = 1120$ 小时对备择假设: $\mu < 1120$ 小时.
- 11.30 某种合金的生产说明书上标明该合金含有 23.2% 铜.从该合金产品中抽取 10 个样品,测得其平均铜含量为 23.5%,标准差为 0.24%.在显著性水平 (a) 0.01, (b) 0.05 下,是否可得结论:产品满足说明书要求?
- 11.31 在问题 11.30 中,在显著性水平 (a) 0.01, (b) 0.05 下检验假设:平均铜含量高于说明中要求.
- 11.32 一效率专家主张在生产线上引进一新型机器,它能较大地减少生产所需时间.因为涉及机器维修费用,管理人员认为,除非生产时间至少能降低 8.0%,否则不值得买机器.6 次合成实验表明生产时间降低了 8.4%,标准差为 0.32%,用显著水平 (a) 0.01, (b) 0.05 下检验假设:应该引进机器.
- 11.33 用商标 A 的汽油,5 辆同样的汽车在相同条件下每加仑平均行驶 22.6 英里,标准差为 0.48 英里.而用商标 B 的汽油,5 辆车平均行驶 21.4 英里,标准差为 0.54 英里.在显著性水平 0.05 下,检验在每加仑汽油所行的平均里程上,商标 A 汽油比商标 B 要好一些.
- 11.34 两种类型的化学溶液 A 和 B,检验它们的 pH 值(即溶液的酸度).分析 A 型的 6 个样品,测得其平均 PH 值为 7.52,标准差为 0.024.分析 B 型的 5 个样品,测得其平均 PH 值为 7.49,标准差为 0.032,在 0.05 显著性水平下,确定两种溶液是否具有不同的 pH 值.
- 11.35 在心理学考试中,一个班的 12 名学生的平均成绩为 78,标准差为 6;而另一个班的 15 名学生的平均成绩为 74,标准差为 8.在显著性水平 0.05 下判断第一组学生是否比第二组学生成绩好?

χ^2 分布

- 11.36 对一自由度为 12 的 χ^2 分布,求 χ^2_c 的值,使得 (a) χ^2_c 右边面积为 0.05, (b) χ^2_c 左边的面积为 0.99, (c) χ^2_c 右边的面积为 0.025.
- 11.37 若自由度 ν 为 (a) 8, (b) 19, (c) 28, (d) 40, 求相应的 χ^2 的临界值,使得 χ^2 分布的右尾面积为 0.05.
- 11.38 若右尾面积为 0.01, 求习题 11.37.
- 11.39 (a) 对自由度为 $\nu = 20$ 的 χ^2 分布,求 χ^2_1 和 χ^2_2 , 使 χ^2_1 与 χ^2_2 之间的面积为 0.95, 这里假设 χ^2_1 右边与 χ^2_2 左边面积相等.
(b) 证明:若 (a) 中不假设 χ^2_2 右边与 χ^2_1 左边面积相等, 则 χ^2_1 与 χ^2_2 的值不惟一.
- 11.40 若变量 U 服从自由度为 $\nu = 7$ 的 χ^2 分布,求 χ^2_1 与 χ^2_2 使得: (a) $P(U > \chi^2_2) = 0.025$, (b) $P(U < \chi^2_1) = 0.50$, (c) $P(\chi^2_1 \leq U \leq \chi^2_2) = 0.90$.
- 11.41 某公司制造的 10 个电灯泡的寿命的标准差为 120 小时,求该公司所有电灯泡寿命的标准差 (a) 95%, (b) 99% 的置信限.
- 11.42 习题 11.41 中若取 25 个电灯泡,其寿命标准差仍为 120 小时,结果会如何?
- 11.43 对自由度 $\nu = 150$, 求 (a) $\chi^2_{0.05}$, (b) $\chi^2_{0.95}$.
- 11.44 对自由度 $\nu = 250$, 求 (a) $\chi^2_{0.025}$, (b) $\chi^2_{0.975}$.
- 11.45 证明:当 ν 值很大时, χ^2_p 可以近似为 $\nu + z_p \sqrt{2\nu}$, 其中 z_p 是标准正态分布的下 p 分位数.
- 11.46 若 100 个电灯泡寿命标准差为 120 小时,用 χ^2 分布求习题 11.41,并将该结论与用第九章中方法所得结论相比较.
- 11.47 习题 11.44 中具有最小宽度的 95% 置信区间是什么?
- 11.48 一公司生产的某电缆的抗断强度的标准差为 240 磅,现对这些电缆的生产程序做一些变动,测得 8 个电缆样品的抗断强度为 300 磅,在显著性水平 (a) 0.05, (b) 0.01 下分别检验假设:变动后标准差有明显增加.

- 11.49 一城市在过去 100 年中年温度的标准差为 16 华氏度,现就最近 15 年中取每月第 15 天的温度得平均温度,并求得其年温度标准差为 10 华氏度.在显著性水平(a) 0.05, (b) 0.01 下检验假设:该城市现在温度比过去变化要小些.

F 分布

- 11.50 对以下每种情形求 F 值:
(a) $\nu_1 = 8, \nu_2 = 10$, 求 $F_{0.95}$; (c) $N_1 = 16, N_2 = 25$, 求 $F_{0.95}$;
(b) $\nu_1 = 24, \nu_2 = 11$, 求 $F_{0.99}$; (d) $N_1 = 21, N_2 = 23$, 求 $F_{0.99}$.
- 11.51 对 $\nu_1 = 22, \nu_2 = 27$, 求 $F_{0.95}$.
- 11.52 从两个服从正态分布具有方差 40 和 60 的总体中分别抽取样本容量为 10 和 15 的样本,若样本方差为 90 和 50,试确定在显著性水平(a) 0.05, (b) 0.01 下,样本 1 方差是否比样本 2 方差明显偏大?
- 11.53 两个灯泡公司 A 和 B,其灯泡寿命都近似服从正态分布,且标准差分别为 20 小时和 27 小时,若从 A 公司选 16 个灯泡, B 公司选 20 个灯泡,测得其标准差分别为 15 小时和 40 小时,则在显著性水平(a) 0.05, (b) 0.01 下,是否可得结论:公司 A 的灯泡变异性比公司 B 的要小?

第十二章 χ^2 检 验

观察频数和理论频数

正如我们所知,从样本中所得结论往往与期望中的理论结论是不相符的.例如,我们把一枚均匀硬币抛 100 次,理论上讲应该出现正面 50 次,反面 50 次,但事实上这种结论是很少出现的.

假如在一特殊样本中,一事件集合 $E_1, E_2, E_3, \dots, E_k$ (见表 12.1) 通过观察知其发生频数分别为 $o_1, o_2, o_3, \dots, o_k$, 称之为**观察频数**, 而根据概率规律, 它们期望发生的频数应该分别是 $e_1, e_2, e_3, \dots, e_k$, 称之为**理论频数**或**期望频数**. 通常, 我们希望知道这些观察频数是否显著不同于它们的理论频数.

表 12.1

事件	E_1	E_2	E_3	\dots	E_k
观察频数	o_1	o_2	o_3	\dots	o_k
理论频数	e_1	e_2	e_3	\dots	e_k

χ^2 的定义

存在于观察频数与理论频数之间的**差异**由统计量 χ^2 来度量, 其中 χ^2 由下式给出:

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \dots + \frac{(o_k - e_k)^2}{e_k} = \sum_{j=1}^k \frac{(o_j - e_j)^2}{e_j} \quad (1)$$

若设总频数为 N , 则

$$\sum o_j = \sum e_j = N \quad (2)$$

另一等价于(1)式的表达式为(见习题 12.11)

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N \quad (3)$$

若 $\chi^2 = 0$, 则观察频数和理论频数完全相同; 若 $\chi^2 > 0$, 则它们不相同, 且随着 χ^2 值的增大, 两者之间的差距也越大.

当理论频数大于或等于 5 时, χ^2 的抽样分布非常近似于第十一章中曾提过的用

$$Y = Y_0(\chi^2)^{\frac{1}{2}(\nu-2)} e^{-\frac{1}{2}\chi^2} = Y_0\chi^{\nu-2} e^{-\frac{1}{2}\chi^2} \quad (4)$$

来表示的 χ^2 分布, 且当理论频数值越大时, 近似程度越好.

自由度 ν 分下列两种情况给出:

- (1) $\nu = k - 1$, 当总体参数无需用样本统计量来估计, 而可直接计算理论频数时;
- (2) $\nu = k - 1 - m$, 当理论频数中有 m 个总体参数要用样本统计量来估计方可计算时.

显著性检验

在实际中, 理论频数是基于假设 H_0 而进行计算的. 如果在假设下, 由(1)或(3)式计算出的 χ^2 值比某些临界值要大(假如分别代表在显著性水平为 0.05 和 0.01 时的临界值 $\chi_{0.95}^2$ 和 $\chi_{0.99}^2$), 我们将得出结论: 观察频数显著不同于理论频数, 从而在相应显著水平下拒绝假设 H_0 , 否则, 我们将接受它(或至少不拒绝它). 这个过程, 我们称之为假设或显著性的 χ^2 检验.

当 χ^2 的值非常接近 0 时, 我们必须特别注意, 因为观察频数与理论频数值一致这种情况

毕竟是很少的. 为了估计这种情形, 我们首先确定 χ^2 的值是否小于 $\chi_{0.05}^2$ (或是 $\chi_{0.01}^2$), 由此决定两频数在显著性水平 0.05 (或 0.01) 下的一致性是否太好.

拟合优度的 χ^2 检验

χ^2 检验能够被用来确定经典分布 (如正态分布和二项分布) 是如何很好地拟合经验分布 (即从样本数据中所得的分布) (见习题 12.12, 12.13).

列联表

表 12.1 中, 观察频数占用了一单行, 被称之为**单向分类表**. 因为列数为 k , 故也称之为 $1 \times k$ 表. 类似可得**双向分类表**, 或 $h \times k$ 表, 在此表中, 观察频数占用了 h 行 k 列, 这样的表通常称之为**列联表**.

相应于 $h \times k$ 列联表中的每个观察频数, 都有一个根据概率规律在某一假设条件下计算得的理论频数与之对应. 这些频数, 占用了列联表中的所有单元, 因此也称之为**单元频数**. 每一行或每一列中的总频数称为**边缘频数**.

为讨论观察频数与理论频数的一致性, 我们计算统计量

$$\chi^2 = \sum_j \frac{(o_j - e_j)^2}{e_j} \quad (5)$$

其中和式包括了列联表中的所有单元. 记号 o_j 和 e_j 分别代表了第 j 单元的观察频数和理论频数. 这和式类似于 (1) 式, 包括 hk 项. 所有观察频数之和 (记为 N) 等于所有理论频数之和 (与 (2) 式比较).

同前, 若理论频数不太小, 则 (5) 式统计量的抽样分布非常近似于由 (4) 式所给出的 χ^2 分布. 这个 χ^2 分布的自由度 ν 分下面两情况给出, 其中 $h > 1$, $k > 1$:

1. $\nu = (h - 1)(k - 1)$, 若理论频数无需通过样本统计量来估计总体参数, 而可直接计算得出 (见习题 12.18).

2. $\nu = (h - 1)(k - 1) - m$, 若理论频数只有通过样本统计量估计出 m 个总体参数后, 方可计算得出.

$h \times k$ 表的显著性检验相似于 $1 \times k$ 表的显著性检验. 理论频数在一特定假设 H_0 下能求得. 通常的假设是两个类别间相互独立.

列联表可推广到更多维数的情形. 例如, 我们能有 $h \times k \times l$ 表, 它代表了三个类别.

关于连续性的 Yates 修正

当连续型分布的结果要应用于离散型分布时, 正如我们在前面章节中所看到的, 必须对连续性做某些修正. 当要用到 χ^2 分布时, 也有类似的修正. (1) 式中的 χ^2 修正为

$$\chi^2(\text{修正}) = \frac{(|o_1 - e_1| - 0.5)^2}{e_1} + \frac{(|o_2 - e_2| - 0.5)^2}{e_2} + \dots + \frac{(|o_k - e_k| - 0.5)^2}{e_k} \quad (6)$$

这种修正通常被称为 **Yates 修正**. 对 (5) 式中 χ^2 也有类似的修正.

一般情况下, 只有自由度 $\nu = 1$ 时才做修正. 对大样本情形, 修正或非修正的 χ^2 所得结论是一样的, 当然在临界值处会有点麻烦 (见习题 12.8). 对于理论频数值在 5 到 10 之间的小样本事件而言, 最好是能比较一下修正前和修正后的 χ^2 所得的结论. 如果对某一假设, 所得结论一致, 例如在 0.05 显著性水平下拒绝假设, 则不会遇到什么麻烦. 如果他们所得结论不同, 则可以求助于增大样本容量, 若这种方法不适用, 则可运用概率的方法来解决, 包括第六章中的**多项分布**.

计算 χ^2 的简单公式

当不需要估计总体参数, 由样本频数可直接计算理论频数时, 有一个计算 χ^2 的简单公

式.下面给出了 2×2 和 2×3 列联表的结果.

2×2 表:

$$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{(a_1 + b_1)(a_2 + b_2)(a_1 + a_2)(b_1 + b_2)} = \frac{N\Delta^2}{N_1N_2N_AN_B} \quad (7)$$

表 12.2

	I	II	总数
A	a_1	a_2	N_A
B	b_1	b_2	N_B
总数	N_1	N_2	N

表 12.3

	I	II	III	总数
A	a_1	a_2	a_3	N_A
B	b_1	b_2	b_3	N_B
总数	N_1	N_2	N_3	N

其中

$\Delta = a_1b_2 - a_2b_1$, $N = a_1 + a_2 + b_1 + b_2$, $N_1 = a_1 + b_1$, $N_2 = a_2 + b_2$, $N_A = a_1 + a_2$, $N_B = b_1 + b_2$.
(见习题 12.19)

用 Yates 修正, (7) 式变为

$$\chi^2(\text{修正}) = \frac{N \left(|a_1b_2 - a_2b_1| - \frac{1}{2}N \right)^2}{(a_1 + b_1)(a_2 + b_2)(a_1 + a_2)(b_1 + b_2)} = \frac{N \left(|\Delta| - \frac{1}{2}N \right)^2}{N_1N_2N_AN_B} \quad (8)$$

2×3 表:

$$\chi^2 = \frac{N}{N_A} \left[\frac{a_1^2}{N_1} + \frac{a_2^2}{N_2} + \frac{a_3^2}{N_3} \right] + \frac{N}{N_B} \left[\frac{b_1^2}{N_1} + \frac{b_2^2}{N_2} + \frac{b_3^2}{N_3} \right] - N \quad (9)$$

其中我们使用了对所有列联表都有效的一般结论(见习题 12.43):

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N \quad (10)$$

结论(9)能推广到所有的 $2 \times k$ 表, 其中 $k > 3$ (见习题 12.46).

列联系数

列联表中各类别间的独立性或相关程度的度量由下式给出:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \quad (11)$$

C 被称之为**列联系数**. C 值越大, 表明相关程度也越大. 列联表中的行数和列数决定了 C 的最大值, 此值小于或等于 1. 如果列联表中的行数和列数均等于 k , 则 C 的最大值为 $\sqrt{(k-1)/k}$ (见习题 12.22, 12.52, 12.53).

属性相关

因为列联表中各类别通常描绘了个体或事物的特性, 我们通常称之为**属性**, 其独立或相关的程度则称之为**属性相关**. 对 $k \times k$ 表, 我们定义

$$r = \sqrt{\frac{\chi^2}{N(k-1)}} \quad (12)$$

作为属性间(各类别间)的相关系数, 其中 $0 \leq r \leq 1$ (见习题 12.24). 对 2×2 表, 其相关性通常称之为四项相关.

数值变量的一般相关问题在第十四章中给予讨论.

χ^2 的可加性

假设重复试验所产生的 χ^2 样本值分别为 $\chi_1^2, \chi_2^2, \chi_3^2, \dots$, 其自由度分别为 $\nu_1, \nu_2, \nu_3, \dots$

那么所有这些试验的结果可以看成是一个试验结果,其 χ^2 值为 $\chi_1^2 + \chi_2^2 + \chi_3^2 + \cdots$, 其自由度为 $\nu_1 + \nu_2 + \nu_3 + \cdots$ (见习题 12.25).

习题及解答

χ^2 检验

- 12.1 把一硬币抛 200 次,得 115 次正面和 85 次反面.在显著性水平(a) 0.05, (b) 0.01 下检验假设:硬币是均匀的.

解 正面与反面的观测频数分别为 $o_1 = 115$ 和 $o_2 = 85$, 而理论频数(若硬币是均匀的) $e_1 = 100$ 和 $e_2 = 100$. 因此

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} = \frac{(115 - 100)^2}{100} + \frac{(85 - 100)^2}{100} = 4.50$$

因为分类数为 $k = 2$, $\nu = k - 1 = 2 - 1 = 1$.

(a) 自由度为 1 的临界值 $\chi_{0.05}^2 = 3.84$, 因为 $4.50 > 3.84$, 故我们在显著性水平 0.05 下拒绝硬币是均匀的这个假设.

(b) 自由度为 1 的临界值 $\chi_{0.01}^2 = 6.63$, 因为 $4.50 < 6.63$, 故我们在显著性水平 0.01 下接受硬币是均匀的这个假设.

我们得到结论:观察结果是可能显著的,即硬币可能不均匀.

- 12.2 用 Yates 修正求习题 12.1.

解

$$\begin{aligned}\chi^2(\text{修正}) &= \frac{(|o_1 - e_1| - 0.5)^2}{e_1} + \frac{(|o_2 - e_2| - 0.5)^2}{e_2} \\ &= \frac{(|115 - 100| - 0.5)^2}{100} + \frac{(|85 - 100| - 0.5)^2}{100} \\ &= \frac{14.5^2}{100} + \frac{14.5^2}{100} = 4.205\end{aligned}$$

因为 $4.205 > 3.84$ 且 $4.205 < 6.63$, 故 12.1 所得结论是有效的.

- 12.3 通过用正态分布近似二项分布来求习题 12.1.

解 假设硬币是均匀的,则硬币抛 200 次出现正面次数的均值和标准差应分别为 $\mu = Np = 200 \times 0.5 = 100$, $\sigma = \sqrt{Npq} = \sqrt{200 \times 0.5 \times 0.5} = 7.07$

解法一

$$115 \text{ 次正面的标准值} = \frac{115 - 100}{7.07} = 2.12$$

用双边检验,在显著性水平 0.05 下,若 z 值在区间 $(-1.96, 1.96)$ 之外,我们将拒绝假设:硬币是均匀的.在显著性水平 0.01 下,相应区间为 $(-2.58, 2.58)$.故(如习题 12.1 所得)可得:在 0.05 水平下,我们拒绝假设,但在 0.01 水平下,接受假设.

注意,上面标准值的平方即, $2.12^2 = 4.50$ 等于习题 12.1 中所得的 χ^2 值.这是分类数 $k = 2$ 的 χ^2 检验中必有的结论(见习题 12.10).

解法二 用连续修正,115 或更多次正面等于 114.5 或更多次正面.则 114.5 次正面的标准值 = $(114.5 - 100)/7.07 = 2.05$.这样所得结论与第一种方法相同.

注意,这标准值的平方即 $2.05^2 = 4.20$ 与习题 12.2 中用 Yates 修正所得 χ^2 的修正值一致,这是用 Yates 修正分类数 $k = 2$ 的 χ^2 检验中必有的结论.

- 12.4 表 12.4 给出了一骰子抛掷 120 次所得的观测频数与理论频数.在显著性水平 0.05 下检验假设:骰子是均匀的.

表 12.4

骰子面	1	2	3	4	5	6
观察频数	25	17	15	23	24	16
理论频数	20	20	20	20	20	20

解

$$\begin{aligned}
 \chi^2 &= \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \frac{(o_3 - e_3)^2}{e_3} + \frac{(o_4 - e_4)^2}{e_4} + \frac{(o_5 - e_5)^2}{e_5} + \frac{(o_6 - e_6)^2}{e_6} \\
 &= \frac{(25 - 20)^2}{20} + \frac{(17 - 20)^2}{20} + \frac{(15 - 20)^2}{20} + \frac{(23 - 20)^2}{20} + \frac{(24 - 20)^2}{20} + \frac{(16 - 20)^2}{20} \\
 &= 5.00
 \end{aligned}$$

因为分类数 $k=6$, $\nu=k-1=5$. 对自由度为 5 的临界值 $\chi_{0.95}^2=11.1$. 因为 $5.00 < 11.1$, 故不能拒绝假设: 骰子是均匀的.

对自由度 5, $\chi_{0.05}^2=11.15 < 5.00$. 因此这个一致性还没有好到足以产生怀疑的程度.

- 12.5 表 12.5 给出了某随机数表中的 250 个数中数 0, 1, ..., 9 的分布情况, 问观察频数是否明显不同于理论频数?

表 12.5

数字	0	1	2	3	4	5	6	7	8	9
观察频数	17	31	29	18	14	20	35	30	20	36
理论频数	25	25	25	25	25	25	25	25	25	25

解

$$\begin{aligned}
 \chi^2 &= \frac{(17 - 25)^2}{25} + \frac{(31 - 25)^2}{25} + \frac{(29 - 25)^2}{25} + \frac{(18 - 25)^2}{25} \\
 &\quad + \dots + \frac{(36 - 25)^2}{25} = 23.3
 \end{aligned}$$

对 $\nu=k-1=9$, 临界值 $\chi_{0.99}^2=21.7 < 23.3$, 因此, 在显著性水平 0.01 下, 观察频数明显不同于理论频数.

- 12.6 在 Gregor Mendel 的豌豆实验中, 他观察到有 315 粒圆黄的, 108 粒圆绿的, 101 粒皱黄的, 32 粒皱绿的. 根据他的理论, 这些种类的数量比应是 9:3:3:1, 问在显著性水平 (a) 0.01, (b) 0.05 下, 有无根据对此理论表示怀疑?

解 豌豆的总数量是 $315 + 108 + 101 + 32 = 556$. 按比例 9:3:3:1, 我们期望得到:

$$\begin{aligned}
 \frac{9}{16} \times 556 &= 312.75 \text{ 粒圆黄的}; \frac{3}{16} \times 556 = 104.25 \text{ 粒皱黄的}; \\
 \frac{3}{16} \times 556 &= 104.25 \text{ 粒圆绿的}; \frac{1}{16} \times 556 = 34.75 \text{ 粒皱绿的};
 \end{aligned}$$

因此

$$\begin{aligned}
 \chi^2 &= \frac{(315 - 312.75)^2}{312.75} + \frac{(108 - 104.25)^2}{104.25} + \frac{(101 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} \\
 &= 0.470
 \end{aligned}$$

因为分类数 $k=4$, $\nu=k-1=3$.

(a) 对 $\nu=3$, $\chi_{0.99}^2=11.3$. 因此在 0.01 显著性水平下不能拒绝理论.

(b) 对 $\nu=3$, $\chi_{0.95}^2=7.81$. 因此在 0.05 显著性水平下不能拒绝理论.

得结论: 理论和实验是一致的.

注意, 对 $\nu=3$, $\chi_{0.05}^2=0.352 < 0.470$. 因此, 尽管得到好的一致性, 但得到的结果还是在合理的抽样误差之下.

- 12.7 一罐子里装有大量 4 种不同颜色的弹子: 红的、桔黄的、黄的、绿的. 现随机从里面抓出

12 粒弹子,其中有 2 粒红的,5 粒桔黄的,4 粒黄的,1 粒绿的.检验假设:罐子里含有相同比例的不同颜色弹子.

解 **解法一** 假设罐子里各种弹子所占比例相同,则取出 12 粒弹子,各种颜色的应各占 3 粒.因为理论频数比 5 小,则用 χ^2 近似不太适合.为避免这种情况,我们采用组合分类使理论频数至少为 5.

如果我们拒绝假设,则应采用最好的拒绝假设的组合分类方法.这里,我们把“红的或绿的”分为一类,“桔黄的或黄的”分为一类.则其相应样本值为 3 或 9,而理论频数均为 6.则有

$$\chi^2 = \frac{(3-6)^2}{6} + \frac{(9-6)^2}{6} = 3$$

对 $\nu = 2 - 1 = 1$, $\chi_{0.95}^2 = 3.84$. 因此,我们在显著性水平 0.05 下,不能拒绝假设(尽管在 0.01 水平下,我们将拒绝假设).可以想象,即便各颜色弹子具有相同比例,但观测结果还得依赖偶然性.

解法二 用 Yates 修正,我们有

$$\chi^2(\text{修正}) = \frac{(|3-6|-0.5)^2}{6} + \frac{(|9-6|-0.5)^2}{6} = \frac{2.5^2}{6} + \frac{2.5^2}{6} = 2.1$$

所得结论与上面一致.因为 Yates 修正往往减少 χ^2 值,故上述的结论是当然的.

应该注意,尽管理论频数很小,我们仍用 χ^2 近似,将得

$$\chi^2 = \frac{(2-3)^2}{3} + \frac{(5-3)^2}{3} + \frac{(4-3)^2}{3} + \frac{(1-3)^2}{3} = 3.33$$

对 $\nu = 3$, $\chi_{0.95}^2 = 7.81$. 从而得与上相同结论.但是事实上,对小频数而言, χ^2 近似是不合理的,若组合频数不可取的话,我们必须采用第六章中精确的概率方法.

12.8 把一对骰子抛掷 360 次,得 74 次 7, 24 次 11, 在 0.05 显著性水平下检验假设:这对骰子是均匀的.

解 一对骰子抛掷有 36 种结果,其中得 7 值的有 6 种,得 11 值的有 2 种.故 $P(7) = \frac{6}{36} = \frac{1}{6}$, $P(11) = \frac{2}{36} = \frac{1}{18}$. 因此在 360 次抛掷中,我们期望有 $\frac{1}{6} \times 360 = 60$ 次 7 值和 $\frac{1}{18} \times 360 = 20$ 次 11 值. 则

$$\chi^2 = \frac{(74-60)^2}{60} + \frac{(24-20)^2}{20} = 4.07$$

对 $\nu = 2 - 1 = 1$, $\chi_{0.95}^2 = 3.84 < 4.07$, 故拒绝假设:骰子是均匀的. 然而用 Yates 修正,有

$$\chi^2(\text{修正}) = \frac{(|74-60|-0.5)^2}{60} + \frac{(|24-20|-0.5)^2}{20} = \frac{13.5^2}{60} + \frac{3.5^2}{20} = 3.65$$

因此,基于 χ^2 修正值,在 0.05 显著性水平下,我们不能拒绝假设.

一般来讲,对这里所碰到的大样本事件,用 Yates 修正所得结论更合理一些.但 χ^2 修正值与临界值如此相近,我们不得不犹豫到底取哪一个结论.在这种情况下,如果我们一定要在 0.05 显著性水平下作检验,那么最好的办法是增加样本观测值的个数.否则,如果我们愿意的,可在其他水平(例如 0.10 水平)下,拒绝该假设.

12.9 对有 5 个孩子的 320 户家庭做统计如表 12.6 所示.试问所得结果与假设:“男孩、女孩的出生比率一样”一致吗?

表 12.6

男女孩数	5 男 0 女	4 男 1 女	3 男 2 女	2 男 3 女	1 男 4 女	0 男 5 女	总数
家庭数	18	56	110	88	40	8	320

解 设 p 为生男孩概率, $q = 1 - p$ 为生女孩的概率. 则“5 个男孩”, “4 男 1 女”, …, “5 个女孩”的概率由二项展开式可得

$$(p + q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5$$

若 $p = q = \frac{1}{2}$, 则有

$$P(5 \text{ 个男孩}) = \left(\frac{1}{2}\right)^5 = \frac{1}{32}; \quad P(2 \text{ 男 } 3 \text{ 女}) = 10 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 = \frac{10}{32}$$

$$P(4 \text{ 男 } 1 \text{ 女}) = 5 \left(\frac{1}{2} \right)^4 \left(\frac{1}{2} \right) = \frac{5}{32}; \quad P(1 \text{ 男 } 4 \text{ 女}) = 5 \left(\frac{1}{2} \right) \left(\frac{1}{2} \right)^4 = \frac{5}{32}$$

$$P(3 \text{ 男 } 2 \text{ 女}) = 10 \left(\frac{1}{2} \right)^3 \left(\frac{1}{2} \right)^2 = \frac{10}{32}; \quad P(5 \text{ 个女孩}) = \left(\frac{1}{2} \right)^5 = \frac{1}{32}$$

则用 320 乘以以上概率即得到有 5, 4, 3, 2, 1, 0 个男孩的家庭数分别为 10, 50, 100, 100, 50, 10. 因此

$$\begin{aligned} \chi^2 &= \frac{(18-10)^2}{10} + \frac{(56-50)^2}{50} + \frac{(110-100)^2}{100} + \frac{(88-100)^2}{100} \\ &\quad + \frac{(40-50)^2}{50} + \frac{(8-10)^2}{10} \\ &= 12.0 \end{aligned}$$

对 $\nu = 6 - 1 = 5$, $\chi_{0.95}^2 = 11.1$, $\chi_{0.99}^2 = 15.1$, 因此我们在水平 0.05 下拒绝假设, 但在 0.01 下不能拒绝假设. 因此得结论: 男孩、女孩的出生比率可能不相等.

- 12.10** 对 500 个人进行统计, 得上周有 155 人从一 VCD 租赁商处租用至少一个 VCD. 用双边检验在水平 $\alpha = 0.05$ 下检验假设: 上周有 25% 的人租用至少一个 VCD. 用标准正态分布和 χ^2 分布进行检验, 证明涉及两种分类的 χ^2 检验等价于第十章中对比例所用的显著性检验.

表 12.7

频数	租 VCD	不租 VCD	总数
观察	155	345	500
理论	125	375	500

解 如果零假设成立, 则 $\mu = Np = 500 \times 0.25 = 125$, $\sigma = \sqrt{Npq} = 9.68$. 计算得检验统计量 $z = (155 - 125)/9.68 = 3.10$. 临界值为 ± 1.96 , 故零假设遭到拒绝.

用 χ^2 分布所得结论, 可从表 12.7 中得

$$\chi^2 = \frac{(155 - 125)^2}{125} + \frac{(345 - 375)^2}{375} = 9.6$$

对自由度为 1 的 $\chi_{0.95}^2 = 3.84 < 9.6$, 因此拒绝零假设. 注意, $3.10^2 = 9.6$, $(\pm 1.96)^2 = 3.84$ 或者说 $z^2 = \chi^2$, 即这两种方法是等价的.

- 12.11** (a) 证明本章的公式(1)可写成 $\chi^2 = \sum \frac{o_i^2}{e_j} - N$.

(b) 用(a)中所得结论检验习题 12.6 中所得 χ^2 值.

解 (a) 由定义

$$\begin{aligned} \chi^2 &= \sum \frac{(o_i - e_j)^2}{e_j} = \sum \left(\frac{o_i^2 - 2o_i e_j + e_j^2}{e_j} \right) \\ &= \sum \frac{o_i^2}{e_j} - 2 \sum o_i + \sum e_j = \sum \frac{o_i^2}{e_j} - 2N + N = \sum \frac{o_i^2}{e_j} - N \end{aligned}$$

其中用了本章中的公式(2).

(b)

$$\chi^2 = \sum \frac{o_i^2}{e_j} - N = \frac{315^2}{312.75} + \frac{108^2}{104.25} + \frac{101^2}{104.25} + \frac{32^2}{34.75} - 556 = 0.470$$

拟合优度

- 12.12** 用 χ^2 检验来确定习题 7.31 中表 7.4 所示数据的拟合优度.

解

$$\begin{aligned} \chi^2 &= \frac{(38 - 33.2)^2}{33.2} + \frac{(144 - 161.9)^2}{161.9} + \frac{(342 - 316.2)^2}{316.2} + \frac{(287 - 308.7)^2}{308.7} \\ &\quad + \frac{(164 - 150.7)^2}{150.7} + \frac{(25 - 29.4)^2}{29.4} \\ &= 7.54 \end{aligned}$$

因为在估计理论频数中所用参数量为 $m=1$ (即二项分布中参数 p), 故 $\nu=k-1-m=6-1-1=4$.

对 $\nu=4$, $\chi_{0.95}^2=9.49$, 因此数据的拟合是好的.

对 $\nu=4$, $\chi_{0.05}^2=0.711$. 因为 $\chi^2=7.54>0.711$, 故该拟合又不是很好到难以相信的程度.

12.13 确定习题 7.33 中表 7.6 所示数据的拟合优度.

解

$$\chi^2 = \frac{(5-4.13)^2}{4.13} + \frac{(18-20.68)^2}{20.68} + \frac{(42-38.92)^2}{38.92} + \frac{(27-27.71)^2}{27.71} + \frac{(8-7.43)^2}{7.43} = 0.959$$

因为在估计理论频数中所用参数量为 $m=2$ (即正态分布中的期望 μ 和标准差 σ), $\nu=k-1-m=5-1-2=2$.

对 $\nu=2$, $\chi_{0.95}^2=5.99$, 故说数据拟合是好的.

对 $\nu=2$, $\chi_{0.05}^2=0.103<0.959$, 故说该拟合又不是“太好”.

列联表

12.14 用 χ^2 检验解习题 10.20.

解 解法一 问题中条件由表 12.8(a) 给出. 在零假设 H_0 : “血清没有效力”下, 我们期望每一组中有 70 人康复, 30 人没有康复, 如表 12.8(b) 所示. 注意, 假设 H_0 即为康复不依赖于血清的使用 (即各类别是相互独立的).

表 12.8(a) 观察频数

	康复	未康复	总数
A 组(用血清)	75	25	100
B 组(未用血清)	65	35	100
总数	140	60	200

表 12.8(b) 在 H_0 下的理论频数

	康复	未康复	总数
A 组(用血清)	70	30	100
B 组(未用血清)	70	30	100
总数	140	60	200

$$\chi^2 = \frac{(75-70)^2}{70} + \frac{(65-70)^2}{70} + \frac{(25-30)^2}{30} + \frac{(35-30)^2}{30} = 2.38$$

为确定自由度值, 考虑表 12.9, 此表与表 12.8 相同, 只是仅写出其中的总数值. 很显然, 我们可以在 4 个空格中任一个空格填上一个数字, 但一旦这个数字填上了, 其他空格处的值由于总数的确定而惟一确定下来了, 故说该处自由度为 1.

表 12.9

	康复	未康复	总数
A 组			100
B 组			100
总数	140	60	200

解法二 由公式(见习题 12.18), $\nu=(h-1)(k-1)=1$. 对自由度为 1, $\chi_{0.95}^2=3.84>2.38$, 因此在 0.05 水平下, 得出结论无效. 即我们在此水平下, 不能拒绝 H_0 . 我们也可以得出结论: 在进一步检验之前, 血清是无效的.

注意, $\chi^2=2.38$ 是 z 值的平方, $z=1.54$ 是在习题 10.20 中所得结论. 一般来讲, 在 2×2 列联表中涉及样本比例的 χ^2 检验等价于用正态近似的比例差异的显著性检验.

同时注意, 用 χ^2 的单边检验等价于用 χ 的双边检验. 例如, $\chi^2>\chi_{0.95}^2$ 相当于 $\chi>\chi_{0.95}$ 或 $\chi<-\chi_{0.95}$. 因为对 2×2 列联表来讲, χ^2 即为 z 的平方, 故可以说 χ 为 z . 因此, 在 0.05 水平下, 用 χ^2 检验而拒绝某假设等价于在水平 0.10 下, 用 z 的双边检验拒绝该假设.

12.15 用 Yates 修正来讨论 12.14.

解

$$\chi^2(\text{修正}) = \frac{(175 - 70 - 0.5)^2}{70} + \frac{(165 - 70 - 0.5)^2}{70} + \frac{(125 - 30 - 0.5)^2}{30} + \frac{(135 - 30 - 0.5)^2}{30} = 1.93$$

因此说习题 12.14 中所得结论是有效的. 事实上这可由 Yates 修正总是减少 χ^2 值而得到.

- 12.16 一电话公司对不同年龄拥有手机的情况进行了调查, 其中对 1000 人调查所得结果如表 12.10 所示. 检验假设 H_0 : 不同年龄层拥有手机的比例是一样的.

表 12.10

手机	18~24	25~54	55~64	≥65	总数
有	50	80	70	50	250
无	200	170	180	200	750
总数	250	250	250	250	1000

解 在 H_0 假设成立下, 每一年龄层拥有手机的比例应是 $250/1000 = 25\%$. 故每一年龄层不拥有手机的百分比为 75% . 理论频数如表 12.11 所示.

χ^2 统计量的计算值由表 12.12 给出.

χ^2 的自由度为 $\nu = (h-1)(k-1) = 3$, $\chi_{0.95}^2 = 7.81$, 而 $14.3 > 7.81$, 故我们拒绝零假设而得结论: 4 个年龄层拥有手机比例是不同的.

表 12.11

手机	18~24	25~54	55~64	≥65	总数
有	250 的 25% = 62.5	250 的 25% = 62.5	250 的 25% = 62.5	250 的 25% = 62.5	250
无	250 的 75% = 187.5	250 的 75% = 187.5	250 的 75% = 187.5	250 的 75% = 187.5	750
总数	250	250	250	250	1000

表 12.12

行, 列	o	e	$(o-e)$	$(o-e)^2$	$(o-e)^2/e$
1, 1	50	62.5	-12.5	156.25	2.5
1, 2	80	62.5	17.5	306.25	4.9
1, 3	70	62.5	7.5	56.25	0.9
1, 4	50	62.5	-12.5	156.25	2.5
2, 1	200	187.5	12.5	156.25	0.8
2, 2	170	187.5	-17.5	306.25	1.6
2, 3	180	187.5	-7.5	56.25	0.3
2, 4	200	187.5	12.5	156.25	0.8
Σ	1000	1000	0		14.3

- 12.17 用 Minitab 解习题 12.16.

解 下面是由 Minitab 软件得到的习题 12.16 之解. 观察频数和理论频数显示在检验统计量的计算中. 注意, 当显著性水平超过 0.002 时零假设就被拒绝.

Data Display

Row	18-24	25-54	55-64	65 or more
1	50	80	70	50
2	200	170	180	200

MTB>chisquare c1-c4

Chi-Square Test

Expected counts are printed below observed counts

	18-24	25-54	55-64	65 or mo	Total
1	50	80	70	50	250
	62.50	62.50	62.50	62.50	
2	200	170	180	200	750
	187.50	187.50	187.50	187.50	
Total	250	250	250	250	1000

Chi-Sq= 2.500 + 4.900 + 0.900 + 2.500 +
0.833 + 1.633 + 0.300 + 0.833 = 14.400

DF=3, P-Value=0.002

12.18 证明对 $h \times k$ 列联表, 其自由度为 $(h-1) \times (k-1)$, 其中 $h > 1$, $k > 1$.

解 在 h 行 k 列的表中, 我们可以对每行每列中的一个数不予考虑, 因为利用总数值是可以得到这些数的. 也就是说对 $h \times k$ 表中, 有 $(h-1) \times (k-1)$ 个数可由我们自由填写, 而剩下的则自动惟一确定下来了. 因此说该表的自由度为 $(h-1) \times (k-1)$.

注意, 在理论频数中所需的总体参数是已知的情况下, 该结论是成立的.

12.19 (a) 证明在表 12.13(a) 中所示的 2×2 列联表中

$$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{N_1N_2N_A N_B}$$

(b) 参照习题 12.14 中数据来阐明(a)的结论.

表 12.13(a) 观察结果

	I	II	总数
A	a_1	a_2	N_A
B	b_1	b_2	N_B
总数	N_1	N_2	N

表 12.13(b) 理论结果

	I	II	总数
A	N_1N_A/N	N_2N_A/N	N_A
B	N_1N_B/N	N_2N_B/N	N_B
总数	N_1	N_2	N

解 如习题 12.14 所示, 在零假设下所期望的结果由表 12.13(b) 给出, 则

$$\chi^2 = \frac{(a_1 - N_1N_A/N)^2}{N_1N_A/N} + \frac{(a_2 - N_2N_A/N)^2}{N_2N_A/N} + \frac{(b_1 - N_1N_B/N)^2}{N_1N_B/N} + \frac{(b_2 - N_2N_B/N)^2}{N_2N_B/N}$$

但

$$a_1 - \frac{N_1N_A}{N} = a_1 - \frac{(a_1 + b_1)(a_1 + a_2)}{a_1 + b_1 + a_2 + b_2} = \frac{a_1b_2 - a_2b_1}{N}$$

同样地

$$a_2 - \frac{N_2N_A}{N} = b_1 - \frac{N_1N_B}{N} = b_2 - \frac{N_2N_B}{N} = \frac{a_1b_2 - a_2b_1}{N}$$

因此我们能写成

$$\chi^2 = \frac{N}{N_1N_A} \left(\frac{a_1b_2 - a_2b_1}{N} \right)^2 + \frac{N}{N_2N_A} \left(\frac{a_1b_2 - a_2b_1}{N} \right)^2 + \frac{N}{N_1N_B} \left(\frac{a_1b_2 - a_2b_1}{N} \right)^2 + \frac{N}{N_2N_B} \left(\frac{a_1b_2 - a_2b_1}{N} \right)^2$$

可化简成

$$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{N_1N_2N_A N_B}$$

(b) 在习题 12.14 中, $a_1 = 75$, $a_2 = 25$, $b_1 = 65$, $b_2 = 35$, $N_1 = 140$, $N_2 = 60$, $N_A = 100$, $N_B = 100$, $N = 200$, 则由(a)所得, 有

$$\chi^2 = \frac{200 \times (75 \times 35 - 25 \times 65)^2}{140 \times 60 \times 100 \times 100} = 2.38$$

用 Yates 修正, 其结果与习题 12.15 中结论相同:

$$\begin{aligned}\chi^2(\text{修正}) &= \frac{N \left(|a_1 b_2 - a_2 b_1| - \frac{1}{2} N \right)^2}{N_1 N_2 N_A N_B} \\ &= \frac{200 \times (|75 \times 35 - 25 \times 65| - 100)^2}{140 \times 60 \times 100 \times 100} = 1.93\end{aligned}$$

- 12.20** 对 900 位男人和 900 位女人调查以便确定他们是否希望更多的联邦机构来帮助照顾小孩, 其中有 40% 女人和 36% 男人对此表示同意. 检验零假设: 表示同意的比例相同, 对应的备择假设为: 表示同意的比例不同. 并说明涉及两样本的比例的 χ^2 检验等价于第十章中用正态近似的比例差异的显著性检验.

解 在假设 H_0 下,

$$\begin{aligned}\mu_{P_1 - P_2} &= 0 \\ \sigma_{P_1 - P_2} &= \sqrt{pq \left(\frac{1}{N_1} + \frac{1}{N_2} \right)} = \sqrt{0.38 \times 0.62 \times \left(\frac{1}{900} + \frac{1}{900} \right)} = 0.0229\end{aligned}$$

其中 p 由两个样本的联合比例来估计, 即

$$p = \frac{360 + 324}{900 + 900} = 0.38, \quad q = 1 - p = 0.62$$

正态近似检验统计量为

$$Z = \frac{P_1 - P_2}{\sigma_{P_1 - P_2}} = \frac{0.40 - 0.36}{0.0229} = 1.7467$$

由 Minitab 软件给出的 χ^2 分析的解如下:

Chi-Square Test

Expected counts are printed below observed counts

	males	females	Total
1	324	360	684
	342.00	342.00	
2	576	549	1116
	558.00	558.00	
Total	900	900	1800

$$\begin{aligned}\text{Chi-Sq} &= 0.947 + 0.947 + \\ &\quad 0.581 + 0.581 = 3.056\end{aligned}$$

$$\text{DF} = 1, \text{P-Value} = 0.080$$

正态检验统计量值的平方是 $1.7467^2 = 3.056$, 正如等于 χ^2 统计量的值. 因此, 两种检验是等价的, 它们的 p -值总是相同的.

列联系数

- 12.21** 对习题 12.14 中列联表所示数据, 求其列联系数.

解

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{2.38}{2.38 + 200}} = \sqrt{0.01176} = 0.1084$$

- 12.22** 对习题 12.14 中列联表求 C 的最大值.

解 当两种类别完全相依时, C 取最大值. 此时, 所有用了血清的人都会康复, 没有用血清的都不会康复得很好. 则列联表将如表 12.14 所示.

表 12.14

	康复	未康复	总数
A 组(用血清)	100	0	100
B 组(未用血清)	0	100	100
总数	100	100	200

在假设完全独立情况下, 每个单元的理论频数均为 50, 故

$$\chi^2 = \frac{(100-50)^2}{50} + \frac{(0-50)^2}{50} + \frac{(0-50)^2}{50} + \frac{(100-50)^2}{50} = 200$$

因此 C 最大值为: $\sqrt{\chi^2/(\chi^2+N)} = 0.7071$.

一般地, 对行与列数均为 k 的列联表中完全相依情形, 非零的单元频数会在从上左到下右的对角线上. 此时, $C_{\max} = \sqrt{(k-1)/k}$. (见习题 12.52 和 12.53)

属性相关

12.23 对习题 12.14 中表 12.8, (a) 不用, (b) 用 Yates 修正来求相关系数.

解 (a) 因为 $\chi^2 = 2.38$, $N = 200$, $k = 2$, 则有

$$r = \sqrt{\frac{\chi^2}{N(k-1)}} = \sqrt{\frac{2.38}{200}} = 0.1091$$

表明用血清与康复之间存在很小的联系.

(b) 从习题 12.15, 得

$$r(\text{修正}) = \sqrt{1.93/200} = 0.0982$$

12.24 证明列联表的相关系数值处于 0 与 1 之间, 如本章(12)式所示.

解 由习题 12.53 知, $\sqrt{\chi^2/(\chi^2+N)}$ 的最大值为 $\sqrt{(k-1)/k}$, 因此

$$\frac{\chi^2}{\chi^2+N} \leq \frac{k-1}{k} \quad k\chi^2 \leq (k-1)(\chi^2+N) \quad k\chi^2 \leq k\chi^2 - \chi^2 + kN - N$$

即

$$k\chi^2 \leq (k-1)N \quad \frac{\chi^2}{N(k-1)} \leq 1 \quad r = \sqrt{\frac{\chi^2}{N(k-1)}} \leq 1$$

而 $\chi^2 \geq 0$, $r \geq 0$, 故有 $0 \leq r \leq 1$.

χ^2 的可加性

12.25 为检验假设 H_0 , 一实验进行了 3 次. 所得 χ^2 值分别为 2.37, 2.86 和 3.54. 每次的自由度均为 1. 证明, 基于任一单个实验, 在 0.05 显著性水平下, 不能拒绝假设 H_0 , 而把三个实验综合起来, 则将拒绝 H_0 .

解 综合 3 次实验所得结果, 根据可加性质, 得 χ^2 的值为 $\chi^2 = 2.37 + 2.86 + 3.54 = 8.77$, 自由度为 $1+1+1=3$. 对 $\nu=3$, $\chi_{0.95}^2 = 7.81 < 8.77$, 故拒绝 H_0 . 而对 $\nu=1$, $\chi_{0.95}^2 = 3.84$, 因此基于一次实验我们不能拒绝 H_0 . 在综合实验中, Yates 修正被略去, 因为它有一个过分正确的趋势.

补充习题

χ^2 检验

- 12.26 一硬币抛掷 60 次得 37 次正面 23 次反面. 在显著性水平 (a) 0.05, (b) 0.01 下检验假设: 硬币是均匀的.
- 12.27 用 Yates 修正讨论习题 12.26.
- 12.28 长期以来, 一批教员对某一特殊课程所给出的成绩平均都是 12% 的 A, 18% 的 B, 40% 的 C, 18% 的 D 和 12% 的 F. 现有一新的教员两个学期以来对这门课程给出了 22 个 A, 34 个 B, 66 个 C, 16 个 D

和 12 个 F . 试在显著性水平 0.05 下, 确定新教员的评分形式是否与其他人一样.

- 12.29 同时抛掷 3 枚硬币 240 次, 其每次正面出现的次数统计如表 12.15 所示, 同时表中也给出了在假设硬币是均匀情况下的期望结果, 试在显著性水平 0.05 下检验该假设.

表 12.15

	0 次正面	1 次正面	2 次正面	3 次正面
观察频数	24	108	95	23
理论频数	30	90	90	30

- 12.30 在某一特定周, 从一公立图书馆所借书的数量如表 12.16 所示, 试在显著性水平 (a) 0.05, (b) 0.01 下检验假设: 所借书的数量与这星期中哪一天无关.

表 12.16

	星期一	星期二	星期三	星期四	星期五
借书数	135	108	120	114	146

- 12.31 一筐子里有 6 个红球和 3 个白球. 从里面任取两个, 记下它们的颜色, 然后放回筐子, 如此反复取 120 次, 所得结果如表 12.17 所示.

(a) 确定理论频数.

(b) 在显著性水平 0.05 下, 确定观察结果与期望结果是否一致.

表 12.17

	0 红 2 白	1 红 1 白	2 红 0 白
取球次数	6	53	61

- 12.32 从 4 台机器所生产的产品中各任取 200 个螺栓, 发现它们的次品数分别为 2, 9, 10 和 3. 试在显著性水平 0.05 下, 确定这些机器是否有明显差异.

拟合优度

- 12.33 (a) 用 χ^2 检验来确定习题 7.75 中表 7.9 的数据的拟合优度.

(b) 在 0.05 显著性水平下讨论该拟合是否很好?

- 12.34 用 χ^2 检验确定 (a) 习题 3.59 中表 3.8, (b) 习题 3.61 中表 3.10 的数据的拟合优度. 并在显著性水平 0.05 下讨论这些拟合是否很好?

- 12.35 用 χ^2 检验确定 (a) 习题 7.79 中表 7.9, (b) 习题 7.80 中表 7.10 的数据的拟合优度. 问 (a) 中所得的结果与习题 12.23 的结论一致吗?

列联表

- 12.36 表 12.18 给出了某实验室接种某种疫苗的效力的一次实验结果, 该实验室专用于研究动物的抗癌性. 试在显著性水平 (a) 0.01, (b) 0.05 下检验假设: 接种与不接种之间没有什么差异 (即接种疫苗与抗癌没有关系).

表 12.18

	患病	未患病
接种疫苗	9	42
未接种疫苗	17	28

表 12.19

	通过	未通过
班 A	72	17
班 B	64	23

- 12.37 用 Yates 修正讨论习题 12.36.

12.38 表 12.19 给出了 A 班与 B 班各通过某一考试的情况. 试在显著性水平(a)0.05, (b) 0.01 下检验假设: 两个班没多大差异. 分别用和不用 Yates 修正来检验该假设.

12.39 有一群病人埋怨他们睡眠不好, 于是让一些人服用安眠药, 另一些人服用糖衣药片(他们都认为自己服的是安眠药). 表 12.20 给出了事后他们的不同反应情况. 假设所有病人所言属实, 则在显著性水平 0.05 下检验假设: 安眠药与糖衣药片之间没有区别.

表 12.20

	睡眠好	睡眠不好
服安眠药	44	10
服糖衣药	81	35

12.40 在一次全国性的重大决策上, 民主派与共和派人士表决如表 12.21 所示. 试在显著性水平(a) 0.01, (b) 0.05 下检验假设: 在有关这次决策问题上, 两党派没有多大差异.

表 12.21

	赞成	反对	未表态
民主派	85	78	37
共和派	118	61	25

12.41 表 12.22 给出了学生的物理成绩和数学成绩情况. 试在(a) 0.01, (b) 0.05 显著性水平下检验假设: 物理成绩与数学成绩是相互独立的.

表 12.22

		数学		
		高等	中等	低等
物理	高等	56	71	12
	中等	47	163	38
	低等	14	42	85

12.42 为确定司机年龄是否会影响到交通事故的发生次数, 对此进行了一项调查, 其结果由表 12.23 给出. 在显著性水平(a) 0.01, (b) 0.05 下检验假设: 事故发生次数与司机年龄无关. 在抽样方法及其他要考虑的原因中有哪些会影响到你的结论?

表 12.23

		司机年龄				
		21~30	31~40	41~50	51~60	61~70
事故数	0	748	821	786	720	672
	1	74	60	51	66	50
	2	31	25	22	16	15
	>2	9	10	6	5	7

12.43 (a) 证明对所有列联表都有 $\chi^2 = \sum(\sigma_j^2/e_j) - N$, 其中 N 为所有单元的总频数.
(b) 用(a)的结论讨论习题 12.41.

12.44 若令 N_i , N_j 分别表示一列联表的第 i 行和第 j 列的理论频数之和(边缘频数), 证明第 i 行第 j 列的单元理论频数为 $N_i N_j / N$, 其中 N 是所有单元的总频数.

12.45 证明本章公式(9). (提示: 用习题 12.43 和 12.44)

12.46 把本章公式(9)的结果推广至 $2 \times k$ 列联表, 其中 $k > 3$.

12.47 证明本章公式(8).

12.48 类似 $h \times k$ 列联表的一些想法, 讨论 $h \times k \times l$ 列联表的情形, 并指出它们可能的用途.

列联系数

12.49 表 12.24 给出了 200 个学生头发与眼睛颜色间的关系.

- (a) 分别用和不用 Yates 的修正求列联系数.
 (b) 将(a)的结论与列联最大系数作比较.

表 12.24

		头发颜色	
		浅黄色	非浅黄色
眼睛颜色	蓝色	49	25
	非蓝色	30	96

- 12.50 分别用和不用 Yates 修正求(a) 习题 12.36, (b) 习题 12.38 中数据的列联系数.
 12.51 求习题 12.41 中数据的列联系数.
 12.52 证明: 3×3 表的最大列联系数约为 $\sqrt{\frac{2}{3}} = 0.8165$.
 12.53 证明: 对 $k \times k$ 表的最大列联系数为 $\sqrt{(k-1)/k}$.

属性相关

- 12.54 求表 12.24 中数据的相关系数.
 12.55 分别用和不用 Yates 修正求(a) 表 12.18 和(b) 表 12.19 中数据的相关系数.
 12.56 求表 12.22 中数学与物理成绩之间的相关系数.
 12.57 若 C 是 $k \times k$ 表的列联系数, r 是相应的属性相关系数, 证明

$$r = C / \sqrt{(1 - C^2)(k - 1)}.$$

χ^2 的可加性

- 12.58 为检验假设 H_0 , 一实验进行了 5 次, 每次相应自由度均为 4, 所得 χ^2 值分别为 8.3, 9.1, 8.9, 7.8 和 8.6. 证明在 0.05 显著性水平下, 单个考虑每次实验, 不能拒绝 H_0 , 而在 0.005 水平下, 综合实验将拒绝 H_0 .

第十三章 曲线拟合和最小二乘法

变量间的相互关系

通常在实际中,两个(或多个)变量间常存在着一种关系.例如,一成年人的体重某种程度上依赖于他的身高;圆的周长与它的半径相关;一种给定气体的压强与它的温度、体积有关.

通常用数学式子来表述变量间的这些关系是很必要的.

曲线拟合

为确定相关联变量间的一个方程,首先第一步应收集一些数据,这些数据也就是考虑变量的相应值.例如,假设以 X 和 Y 分别表示成人的身高和体重,则 N 个个体的样本将显示其身高 X_1, X_2, \dots, X_N 及相应的体重 Y_1, Y_2, \dots, Y_N .

第二步则是在一直角坐标系上描出这些点: $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$. 这一组点集有时称为**散点图**.

从这个散点图,通常可以观测到一条接近数据的光滑曲线,这样的曲线称为**近似曲线**.例如在图 13-1 中,数据近似一条直线,因此我们说变量间具有线性关系.然而在图 13-2 中,尽管变量间也存在一种关系,但不是线性关系,因此我们称之为非线性关系.

这种求适合所给数据的近似曲线方程的一般问题称为**曲线拟合**.

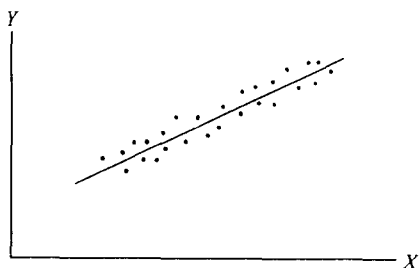


图 13-1

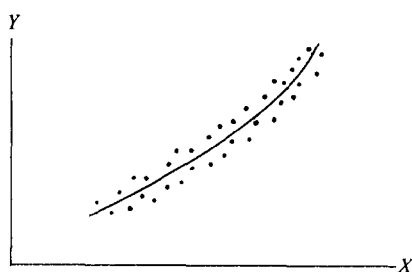


图 13-2

近似曲线的方程

一些一般类型的近似曲线及其方程在下面列出以供参考.除 X 和 Y 外的其他字母均表示常数.变量 X 和 Y 分别表示**自变量**和**因变量**.

$$\text{直线 } Y = a_0 + a_1 X \quad (1)$$

$$\text{抛物线或二次曲线 } Y = a_0 + a_1 X + a_2 X^2 \quad (2)$$

$$\text{三次曲线 } Y = a_0 + a_1 X + a_2 X^2 + a_3 X^3 \quad (3)$$

$$\text{四次曲线 } Y = a_0 + a_1 X + a_2 X^2 + a_3 X^3 + a_4 X^4 \quad (4)$$

$$n \text{ 次曲线 } Y = a_0 + a_1 X + a_2 X^2 + \dots + a_n X^n \quad (5)$$

上述方程的右边分别称为一次、二次、三次、四次、 n 次**多项式**.有时,前 4 个方程所定义的函数分别称为**线性函数**、**二次函数**、**三次函数**和**四次函数**.

下面是在实际中经常用到的一些其他的方程:

$$\text{双曲线 } Y = \frac{1}{a_0 + a_1 X} \text{ 或 } \frac{1}{Y} = a_0 + a_1 X \quad (6)$$

$$\text{指数曲线 } Y = ab^X \text{ 或 } \log Y = \log a + (\log b) X = a_0 + a_1 X \quad (7)$$

$$\text{几何曲线 } Y = aX^b \text{ 或 } \log Y = \log a + b(\log X) \quad (8)$$

$$\text{修正指数曲线 } Y = ab^X + g \quad (9)$$

$$\text{修正几何曲线 } Y = aX^b + g \quad (10)$$

$$\text{古姆波茨曲线 } Y = pq^{b^X} \text{ 或 } \log Y = \log p + b^X(\log q) = ab^X + g \quad (11)$$

$$\text{修正古姆波茨曲线 } Y = pq^{b^X} + h \quad (12)$$

$$\text{逻辑斯谛曲线 } Y = \frac{1}{ab^X + g} \text{ 或 } \frac{1}{Y} = ab^X + g \quad (13)$$

$$Y = a_0 + a_1(\log X) + a_2(\log X)^2 \quad (14)$$

为了确定哪一种曲线适合, 最好能获得变换变量的散点图. 例如, 若 $\log Y$ 与 X 的关系散点图显示线性关系, 则应有(7)式的形式, 若 $\log Y$ 与 $\log X$ 的关系散点图显示线性关系, 则应有(8)式的形式. 有时为了更容易地确定采用哪一种曲线, 通常要用一种特殊图纸. 具有单对数测度尺的图纸称为**半对数图解纸**, 具有双对数测度尺的图纸称为**对数-对数图解纸**.

曲线拟合的徒手法

通常由个人的直观判断来画一条拟合一组数据的近似曲线, 这种方法称为**曲线拟合的徒手法**. 如果曲线方程形式是已知的, 那么可通过选择与方程中心常数个数同样多的点来求出该方程的常数. 例如, 如果曲线是一直线, 则需两个点; 如果曲线是抛物线, 则三个点足够了. 这种方法也有不利之处, 即不同的观察者将得出不同的曲线和方程.

直线

近似曲线的最简单形式就是直线, 其方程可写为

$$Y = a_0 + a_1X \quad (15)$$

给定一直线的任意两点 (X_1, Y_1) , 和 (X_2, Y_2) , 常数 a_0, a_1 便能确定. 直线的最终方程可写为

$$Y - Y_1 = \left(\frac{Y_2 - Y_1}{X_2 - X_1} \right) (X - X_1) \text{ 或 } Y - Y_1 = m(X - X_1) \quad (16)$$

其中

$$m = \frac{Y_2 - Y_1}{X_2 - X_1}$$

称为直线的**斜率**, 代表 Y 的变化量与相应 X 的变化量的商.

当方程写成(15)式的形式时, 常数 a_1 即为斜率 m . 常数 a_0 即为当 $X=0$ 时 Y 的值, 称为 Y 的**截距**.

最小二乘法

为避免在构造直线, 抛物线或其他一些拟合数据的近似曲线中个人判断的不同, 有必要在“最佳拟合直线”、“最佳拟合抛物线”等定义上取得一致.

考虑图 13-3, 其数据点由 $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$ 给出. 对某一给定的 X 值如 X_1 , Y_1 与 X_1 所对应的曲线上的值必存在一定的差异, 如图中所示, 我们记这差异为 D_1 , 有时我们称之为**偏差**、**误差**或**残差**, 其值可能为正, 为负或为 0. 同样地, 相应于值 X_2, \dots, X_N , 我们将获得 D_2, \dots, D_N .

对一组给定的数据拟合一条近似曲线 C , 其“拟合优度”可用数值 $D_1^2 + D_2^2 + \dots + D_N^2$ 来度量. 该数值小, 表示拟合得好; 该数值大, 表示拟合得差. 因此我们有以下定义.

定义 在一组给定数据的所有拟合曲线中, 若某曲线使得 $D_1^2 + D_2^2 + \dots + D_N^2$ 达到最小, 则称该曲线为**最佳拟合曲线**.

使偏差平方和 $D_1^2 + D_2^2 + \dots + D_N^2$ 达到最小这一要求称为**最小二乘原理**, 故最佳拟合曲线

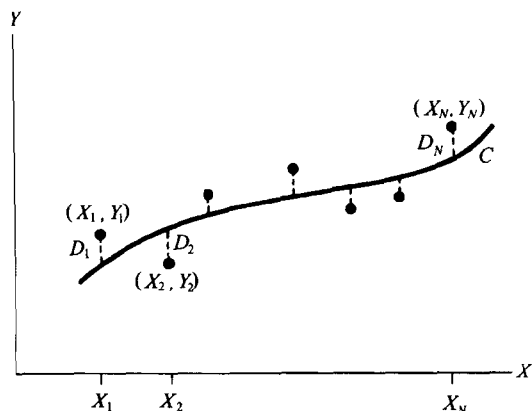


图 13-3

也称为**最小二乘曲线**.特别地,具有这种性质的直线称为**最小二乘直线**,具有这种性质的抛物线称为**最小二乘抛物线**等.

当 X 是自变量, Y 是因变量时,通常都用上面的定义.若 X 是因变量,则上述定义需进行修改,不是考虑竖直偏差,而是考虑水平偏差,也相当于是变换一下 X, Y 轴.这两种定义一般都会导致不同的最小二乘曲线.除非特别指明,以下我们都将 X 作为自变量, Y 作为因变量.

也可以通过考虑每一个数据点到曲线的垂直距离而非竖直距离或水平距离来定义最小二乘曲线.但我们一般不用这一定义.

最小二乘直线

接近一系列点 $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$ 的最小二乘直线具有方程形式:

$$Y = a_0 + a_1 X \quad (17)$$

其中常数 a_0, a_1 通过同时解下面的方程来确定:

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned} \quad (18)$$

这两个方程称为**最小二乘直线(17)的正规方程**.方程(18)中的常数 a_0, a_1 可以由以下公式求得:

$$\begin{aligned} a_0 &= \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2} \\ a_1 &= \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} \end{aligned} \quad (19)$$

正规方程(18)可由以下方程记住,第一个方程由(17)式的两边求和而得(即 $\sum Y = \sum (a_0 + a_1 X) = a_0 N + a_1 \sum X$),而第二个方程则是先用 X 去乘(17)式的两边再求和而得(即, $\sum XY = \sum X(a_0 + a_1 X) = a_0 \sum X + a_1 \sum X^2$).注意这不是正规方程的派生,而仅仅是方便记忆的一种方法.另外也要注意在方程(18)、(19)式中,我们用简便记号 $\sum X, \sum XY$ 等来代替了 $\sum_{j=1}^N X_j, \sum_{j=1}^N X_j Y_j$ 等.

找最小二乘直线的工作有时可以通过变换数据为 $x = X - \bar{X}, y = Y - \bar{Y}$ 而得到简化.这时最小二乘直线的方程可以写成(见习题 13.15):

$$y = \left[\frac{\sum xy}{\sum x^2} \right] x \quad \text{或} \quad y = \left[\frac{\sum xY}{\sum x^2} \right] x \quad (20)$$

特别地,如果 X 有 $\sum X = 0$ (即 $\bar{X} = 0$), 则(20)式可写为

$$Y = \bar{Y} + \left[\frac{\sum XY}{\sum X^2} \right] X \quad (21)$$

方程(20)表明当 $x=0$ 时 $y=0$. 因此最小二乘直线通过点 (\bar{X}, \bar{Y}) , 该点称为数据的**质心**或**重心**.

如果变量 X 是因变量而不是自变量, 那么我们改写方程(17)为: $X = b_0 + b_1 Y$. 以上结论, 若把 X 和 Y 交换一下, a_0 与 a_1 分别以 b_0, b_1 代替, 则依旧成立. 但最小二乘直线一般情况下不同于上面所获得的直线(见习题 13.11 和 13.15(d)).

非线性关系

非线性关系有时可通过一适当的变量变换简化为线性关系(见习题 13.21).

最小二乘抛物线

接近一系列点 $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$ 的最小二乘抛物线具有方程:

$$Y = a_0 + a_1 X + a_2 X^2 \quad (22)$$

其中常数 a_0, a_1, a_2 通过同时解下列方程来确定:

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X + a_2 \sum X^2 \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3 \\ \sum X^2 Y &= a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4 \end{aligned} \quad (23)$$

此方程称为**最小二乘抛物线(22)的正规方程**.

方程(23)事实上是由方程(22)的两边分别乘以 $1, X, X^2$, 然后再对两边求和而得, 故很容易记住. 这种技巧可以推广而获得最小二乘三次曲线、最小二乘四次曲线及一般的对应于方程(5)的最小二乘曲线的正规方程.

正如最小二乘直线中的情形, 若选择 X 使之满足 $\sum X = 0$ 或变换变量 $x = X - \bar{X}, y = Y - \bar{Y}$, 则最小二乘曲线也可获得类似(20)式的简化.

回归

通常, 基于一组样本数据, 我们希望通过一给定的变量 X 的值来估计变量 Y 的值. 这个愿望可以通过拟合样本数据的最小二乘曲线来完成. 这个拟合曲线称为 **Y 关于 X 的回归曲线**, 这是因为 Y 是由 X 值估计得到的.

若我们想从给定的 Y 值来估计 X 的值, 我们将用到 **X 关于 Y 的回归曲线**, 事实上也相当于在散点图中转换变量, 使得 X 是因变量而 Y 是自变量. 这也等价于前面所介绍的在最小二乘曲线的定义中用水平偏差代替竖直偏差.

一般来讲, Y 关于 X 的回归直线或曲线不同于 X 关于 Y 的回归直线或曲线.

时间序列的应用

若自变量 X 是时间, 则数据给出不同时间的 Y 的值. 按时间排列的数据称为**时间序列**. 在这里所讨论的 Y 关于 X 的回归直线或曲线通常称为**趋势直线**或**趋势曲线**, 通常被用于估计、预报或预测.

两个以上变量的问题

两个以上变量的问题往往可以类似两个变量的问题来讨论. 例如, 在变量 X, Y, Z 之间可能存在一种可由下式描述的关系:

$$Z = a_0 + a_1 X + a_2 Y \quad (24)$$

这个方程称为**变量** X, Y, Z 之间的**线性方程**.

在三维直角坐标系中, 这个方程表示一个平面. 样本点 $(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \dots, (X_N, Y_N, Z_N)$ 可能不会“散布”在这个平面上, 但又很接近这个平面, 我们称之为**近似平面**.

通过推广最小二乘法, 我们可以得到接近数据的**最小二乘平面**. 如果我们从给定的 X, Y 值来估计 Z 的值, 那么该平面称为 Z 关于 X, Y 的**回归平面**. 对应于方程(24)的最小二乘平面的正规方程由下式给出:

$$\begin{aligned} \sum Z &= a_0 N + a_1 \sum X + a_2 \sum Y \\ \sum XZ &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum XY \\ \sum YZ &= a_0 \sum Y + a_1 \sum XY + a_2 \sum Y^2 \end{aligned} \quad (25)$$

(25)式同样可通过对(24)两边分别乘以 1, X 和 Y , 然后再求和而得.

比(24)式更复杂的方程同样也可以讨论, 这些都称为**回归(曲)面**. 若变量数目超过 3 个, 则直观上将难以讨论, 因为此时我们需要在四维、五维甚至更高维空间中来讨论.

从两个或更多变量来估计某一变量的问题称为**多重回归**, 这些将在第十五章中有更详细的介绍.

习题及解答

直线

- 13.1 (a)构造一直线使其与表 13.1 所给数据接近.
(b)求这条直线的方程.

表 13.1

X	2	3	5	7	9	10
Y	1	3	7	11	15	17

解 (a)把点 $(2, 1), (3, 3), (5, 7), (7, 11), (9, 15)$ 和 $(10, 17)$ 描在一直角坐标系上, 如图 13-4 所示. 显然所有的点都在一条直线上, 因此该直线完全拟合这些数据.

(b)为确定直线的方程

$$Y = a_0 + a_1 X \quad (26)$$

只需给出两个点即可. 选择点 $(2, 1)$ 和 $(3, 3)$ 分别代入(26)式, 则有

$$1 = a_0 + 2a_1 \quad (27)$$

$$3 = a_0 + 3a_1 \quad (28)$$

同时解方程(27), (28)得 $a_0 = -3, a_1 = 2$. 因此所求方程为

$$Y = 2X - 3 \quad \text{或} \quad Y = -3 + 2X$$

同时我们也可证明点 $(5, 7), (7, 11), (9, 15)$ 和 $(10, 17)$ 也满足该方程, 即在该直线上.

- 13.2 在习题 13.1 中, 求(a) $X = 4$ 时, Y 的值, (b) $X = 15$ 时, Y 的值, (c) $X = 0$ 时, Y 的值, (d) $Y = 7.5$ 时, X 的值, (e) $Y = 0$ 时, X 的值, (f) 当 X 增加一个单位时, Y 的增量.

解 把已知值代入方程 $Y = 2X - 3$, 即可得未知值. (a) $Y = 5$, (b) $Y = 27$, (c) $Y = -3$, (d) $X =$

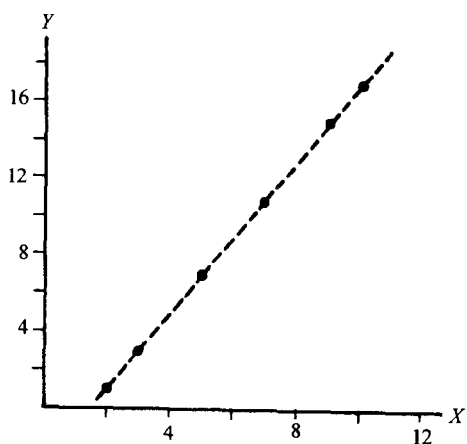


图 13-4

5.25, (e) $X = 1.5$, (f) 当 X 从 2 增加到 3, Y 从 $2 \times 2 - 3 = 1$ 增加到 $2 \times 3 - 3 = 3$, 即当 X 增加一个单位时, Y 相应地增加 2 个单位.

13.3 (a) 证明过点 (X_1, Y_1) 和 (X_2, Y_2) 的直线方程为

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

(b) 求过点 $(2, -3)$ 和 $(4, 5)$ 的直线方程.

解 (a) 直线的一般方程可写为

$$Y = a_0 + a_1 X \quad (29)$$

而 (X_1, Y_1) 和 (X_2, Y_2) 均在直线上, 故满足

$$Y_1 = a_0 + a_1 X_1 \quad (30)$$

$$Y_2 = a_0 + a_1 X_2 \quad (31)$$

(29) 式减 (30) 式得

$$Y - Y_1 = a_1(X - X_1) \quad (32)$$

(31) 式减 (30) 式得

$$Y_2 - Y_1 = a_1(X_2 - X_1) \quad \text{或} \quad a_1 = \frac{Y_2 - Y_1}{X_2 - X_1}$$

把 a_1 的值代入方程 (32), 得

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

证毕.

$\frac{Y_2 - Y_1}{X_2 - X_1}$ 通常简写为 m , 表示 Y 的变化量与相应 X 的变化量的商, 称为直线的斜率. 所求方程可

写为: $Y - Y_1 = m(X - X_1)$.

(b) 解法一 用 (a) 的结论: $X_1 = 2, Y_1 = -3, X_2 = 4, Y_2 = 5$, 因此斜率为

$$m = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{8}{2} = 4$$

故所求方程为: $Y - Y_1 = m(X - X_1)$, 或 $Y - (-3) = 4(X - 2)$, 即, $Y = 4X - 11$.

解法二 用直线的一般方程: $Y = a_0 + a_1 X$, 把点 $(2, -3)$ 与 $(4, 5)$ 代入, 求出 $a_1 = 4, a_0 = -11$, 从而得方程

$$Y = 4X - 11$$

13.4 对习题 13.3(a) 的结论的由来给出一个图像的解释.

解 图 13-5 给出了通过点 $P(X_1, Y_1), Q(X_2, Y_2)$ 的直线, 点 $R(X, Y)$ 表示直线上的任一点.

由三角形 PRT 与三角形 PQS 相似可得

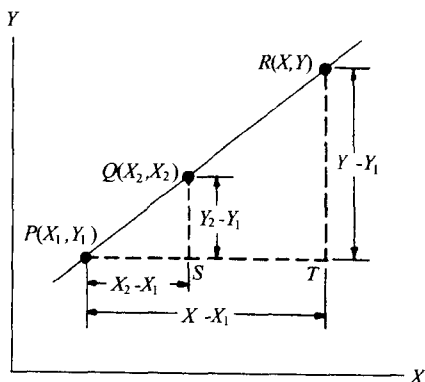


图 13-5

$$\frac{RT}{TP} = \frac{QS}{SP} \quad \text{或} \quad \frac{Y - Y_1}{X - X_1} = \frac{Y_2 - Y_1}{X_2 - X_1} \quad (33)$$

然后用 $(X - X_1)$ 乘以 (33) 式两边, 得

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

即得直线方程.

(33) 式中的比值为斜率 m , 故可写成

$$Y - Y_1 = m(X - X_1)$$

13.5 对过点 $(1, 5)$ 和 $(4, -1)$ 的直线求 (a) 斜率, (b) 方程, (c) Y 的截距, (d) X 的截距.

解 (a) $X_1 = 1, Y_1 = 5, X_2 = 4, Y_2 = -1$, 则斜率

$$m = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{-1 - 5}{4 - 1} = -2$$

负的斜率表明随 X 的增加, Y 减少, 如图 13-6 所示.

(b) 方程为

$$Y - Y_1 = m(X - X_1)$$

即

$$Y - 5 = -2(X - 1)$$

或

$$Y = 7$$

(c) Y 的截距为 $X=0$ 时 Y 的取值. 当 $X=0$ 时

$$Y = 7 - 0 = 7$$

(d) X 的截距为 $Y=0$ 时 X 的取值. 当 $Y=0$ 时

$$0 = 7 - 2X, X = 3.5$$

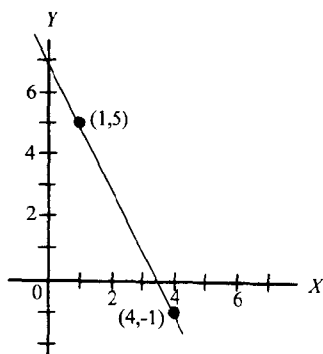


图 13-6

13.6 求平行于直线 $2X + 3Y = 6$ 且过点 $(4, 2)$ 的直线方程.

解 若两直线平行, 则其斜率相等. 从方程 $2X + 3Y = 6$ 得 $Y = 2 - \frac{2}{3}X$. 即得直线的斜率为 $m = -\frac{2}{3}$. 因此所求直线方程为

$$Y - Y_1 = m(X - X_1)$$

即

$$Y - 2 = -\frac{2}{3}(X - 4)$$

或

$$2X + 3Y = 14$$

另解 平行于直线 $2X + 3Y = 6$ 的任一直线方程可写为: $2X + 3Y = c$. 将 $X = 4, Y = 2$ 代入, 即得: $c = 14$. 故所求方程为: $2X + 3Y = 14$.

13.7 求一直线方程, 使其斜率为 -4 , Y 的截距为 16 .

解 在方程 $Y = a_0 + a_1X$ 中, $a_0 = 16$ 为 Y 的截距, $a_1 = -4$ 为斜率, 故所求方程为

$$Y = 16 - 4X$$

13.8 (a) 构造一直线使该直线与表 13.2 中数据接近.

(b) 求该直线方程.

表 13.2

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

解 (a) 把点 $(1, 1), (3, 2), (4, 4), (6, 4), (8, 5), (9, 7), (11, 8)$ 和 $(14, 9)$ 描在一直角坐标系上, 如图 13-7 所示, 即得一接近这些数据的直线. 要找一种排除个人判断影响的方法, 可采用最小二乘法, 见习题 13.11.

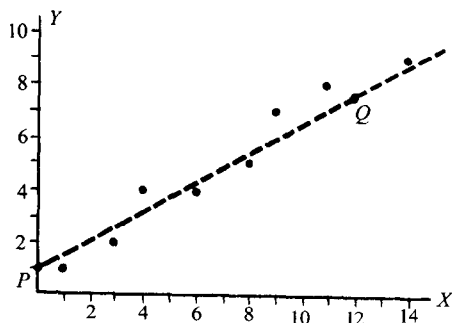


图 13-7

(b)为得到该直线方程,先任选两点,如(0,1)与(12,7.5),代入直线一般方程 $Y = a_0 + a_1x$, 求出 $a_0 = 1, a_1 = 0.542$, 故直线方程为: $Y = 1 + 0.542X$.

13.9 (a)比较近似直线所得 Y 值与表 13.2 中所给的 Y 值.

(b)当 $X = 10$ 时,估计 Y 的值.

解 (a)对 $X = 1, Y = 1 + 0.542 \times 1 = 1.542$ 或 1.5, 对 $X = 3, Y = 1 + 0.542 \times 3 = 2.626$ 或 2.6. 对于 X 的其他值可同样得到相应的 Y 值. 由方程 $Y = 1 + 0.542X$ 所估计的 Y 值, 记为 Y_{est} . 这些估计值与实际值由表 13.3 给出.

表 13.3

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9
Y_{est}	1.5	2.6	3.2	4.3	5.3	5.9	7.0	8.6

(b)当 $X = 10$ 时, $Y = 1 + 0.542 \times 10 = 6.42$ 或 6.4.

13.10 从某一州立大学一年级学生中随机抽取 12 名男生, 测出他们的身高和体重, 如表 13.4 所示.

表 13.4

身高 X (英寸)	70	63	72	60	66	70	74	65	62	67	65	68
体重 Y (磅)	155	150	180	135	156	168	178	160	132	145	139	152

(a)作出这些数据的散点图.

(b)构造一接近数据的直线.

(c)求(b)中所构造直线的方程.

(d)已知某学生身高为 63 英寸, 估计该学生的体重.

(e)已知某学生体重为 168 磅, 估计该学生的身高.

解 (a)散点图如图 13-8 所示.

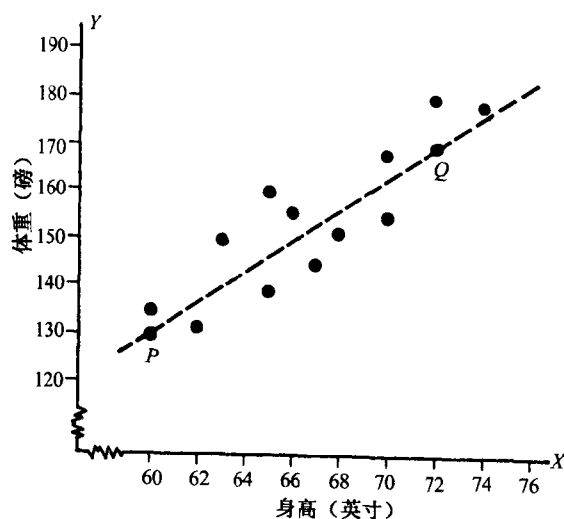


图 13-8

(b)直线亦如图 13-8 所示, 只是所有可能直线中的一条.

(c)选择点 $P(60, 130), Q(72, 170)$, 得 $m = \frac{170 - 130}{72 - 60} = \frac{10}{3}$, 则方程为

$$Y - 130 = \frac{10}{3}(X - 60)$$

即

$$Y = \frac{10}{3}X - 70$$

(d) 若 $X = 63$, 则 $Y = \frac{10}{3} \times 63 - 70 = 140$ 磅.

(e) 若 $Y = 168$, 则 $168 = \frac{10}{3}X - 70$, $X = 71.4$ 或 71 英寸.

最小二乘直线

13.11 对习题 13.8 中数据拟合一最小二乘直线, 其中(a) X 作为自变量, (b) X 作为因变量.

解 (a) 设直线方程为 $Y = a_0 + a_1X$, 其正规方程为

$$\begin{aligned}\sum Y &= a_0N + a_1\sum X \\ \sum XY &= a_0\sum X + a_1\sum X^2\end{aligned}$$

其中相应的求和的值由表 13.5 给出.

表 13.5

X	Y	X^2	XY	Y^2
1	1	1	1	1
3	2	9	6	4
4	4	16	16	16
6	4	36	24	16
8	5	64	40	25
9	7	81	63	49
11	8	121	88	64
14	9	196	126	81
$\sum X = 56$	$\sum Y = 40$	$\sum X^2 = 524$	$\sum XY = 364$	$\sum Y^2 = 256$

因为有 8 对 X, Y 的值, 故 $N = 8$, 则正规方程为

$$8a_0 + 56a_1 = 40$$

$$56a_0 + 524a_1 = 364$$

同时解方程得 $a_0 = \frac{6}{11}$, $a_1 = \frac{7}{11}$, 则所求最小二乘直线为 $Y = \frac{6}{11} + \frac{7}{11}X$.

另解

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} = \frac{40 \times 524 - 56 \times 364}{8 \times 524 - 56^2} = \frac{6}{11} \text{ 或 } 0.545$$

$$a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = \frac{8 \times 364 - 56 \times 40}{8 \times 524 - 56^2} = \frac{7}{11} \text{ 或 } 0.636$$

故所求方程为 $Y = \frac{6}{11} + \frac{7}{11}X$.

(b) 若 X 为因变量, 而 Y 为自变量, 则最小二乘直线为 $X = b_0 + b_1Y$, 其正规方程为

$$\begin{aligned}\sum X &= b_0N + b_1\sum Y \\ \sum XY &= b_0\sum Y + b_1\sum Y^2\end{aligned}$$

由表 13.5, 可求得

$$b_0 = -\frac{1}{2}, b_1 = \frac{3}{2}$$

故所求直线为 $X = -\frac{1}{2} + \frac{3}{2} Y$.

也可类似上面另一法而求得 b_0, b_1 .

注意 (a)与(b)所得的二条最小二乘直线是不同的.

- 13.12** 对习题 13.10 中体重/身高数据, 设身高为自变量, 体重为因变量, 用 Minitab 求最小二乘直线, 并把观察的数据与最小二乘直线上相应的点标在同一坐标系上.

解 Minitab 输出结果如下. 命令 `regress c2 on 1 variable in c1` 显示出最小二乘直线方程, $\text{weight} = -60.7 + 3.22 \text{ height}$. 为了充分欣赏计算软件的能力, 参见寻求最小二乘直线的必要的计算方法, 如习题 13.17 给出的那样.

MTB>print c1 c2

Data Display

Row	height	weight
1	70	155
2	63	150
3	72	180
4	60	135
5	66	156
6	70	168
7	74	178
8	65	160
9	62	132
10	67	145
11	65	139
12	68	152

MTB>regress c2 on 1 variable in c1

Regression Analysis

The regression equation is

weight = - 60.7 + 3.22 height

在图 13-9 中, 观察数据值用方框表示, 最小二乘直线上的相应点用 + 字符号表示.

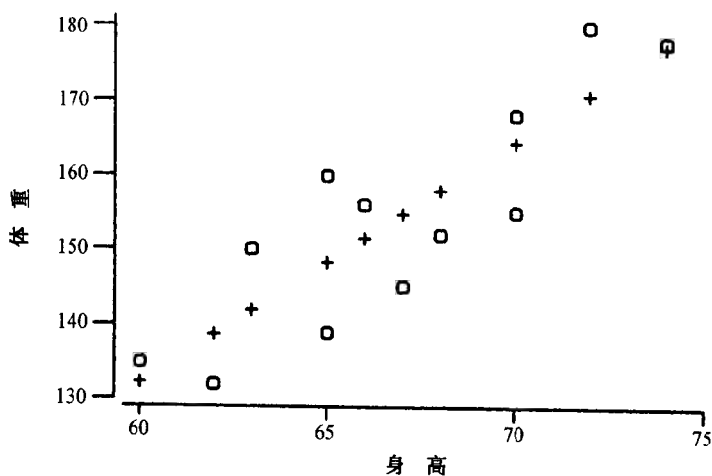


图 13-9

- 13.13** (a)证明习题 13.11 中所得的两条最小二乘直线相交于点 (\bar{X}, \bar{Y}) .
 (b)当 $X=12$ 时, 求 Y 值.
 (c)当 $Y=3$ 时, 求 X 值.

解 $\bar{X} = \frac{\sum X}{N} = \frac{56}{8} = 7, \bar{Y} = \frac{\sum Y}{N} = \frac{40}{8} = 5.$

因此称为质心的点 (\bar{X}, \bar{Y}) 为点 $(7, 5)$.

(a)点 $(7, 5)$ 满足方程: $Y = \frac{6}{11} + \frac{7}{11}X$, 即, $5 = \frac{6}{11} + \frac{7}{11} \times 7$. 点 $(7, 5)$ 也满足方程: $X = -\frac{1}{2} + \frac{3}{2}Y$, 即,

$$7 = -\frac{1}{2} + \frac{3}{2} \times 5.$$

另解 同时解方程:

$$Y = \frac{6}{11} + \frac{7}{11}X \text{ 与 } X = -\frac{1}{2} + \frac{3}{2}Y$$

求得 $X=7, Y=5$, 得相交点为 $(7, 5)$.

(b)将 $X=12$ 代入直线 $Y = \frac{6}{11} + \frac{7}{11}X$, 得 $Y=8.2$.

(c)将 $Y=3$ 代入直线 $X = -\frac{1}{2} + \frac{3}{2}Y$, 得 $X=4.0$.

13.14 证明最小二乘直线总是通过点 (\bar{X}, \bar{Y}) .

解 情形 1 (X 是自变量)最小二乘直线方程为

$$Y = a_0 + a_1X \quad (34)$$

其正规方程之一为

$$\sum Y = a_0N + a_1 \sum X \quad (35)$$

将(35)式等号两边同时除以 N , 则得

$$\bar{Y} = a_0 + a_1\bar{X} \quad (36)$$

(34)式减去(36)式, 得

$$Y - \bar{Y} = a_1(X - \bar{X}) \quad (37)$$

即表明直线通过点 (\bar{X}, \bar{Y}) .

情形 2 (Y 是自变量)同情形 1, 只需交换 X 与 Y 的位置, 同时分别将 a_0, a_1 用 b_0, b_1 代替, 则可得到方程:

$$X - \bar{X} = b_1(Y - \bar{Y}) \quad (38)$$

从而说明直线通过点 (\bar{X}, \bar{Y}) .

注意, 直线(37)与(38)并非一致, 但相交于点 (\bar{X}, \bar{Y}) .

13.15 (a) 设 X 为自变量, 证明最小二乘直线方程可写成

$$y = \left[\frac{\sum xy}{\sum x^2} \right] x \quad \text{或} \quad y = \left[\frac{\sum xY}{\sum x^2} \right] x$$

其中 $x = X - \bar{X}, y = Y - \bar{Y}$.

(b) 若 $\bar{X}=0$, 证明(a)中最小二乘直线方程可写为

$$Y = \bar{Y} + \left[\frac{\sum XY}{\sum X^2} \right] X$$

(c) 设 Y 为自变量, 求对应(a)中的最小二乘直线方程.

(d) 证明(a)与(c)中直线未必一样.

解 (a) 方程(37)可写成 $y = a_1x$, 其中 $x = X - \bar{X}, y = Y - \bar{Y}$.

由正规方程(18)解得((19)式中的第二式)

$$\begin{aligned} a_1 &= \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} \\ &= \frac{N \sum (x + \bar{X})(y + \bar{Y}) - [\sum (x + \bar{X})][\sum (y + \bar{Y})]}{N \sum (x + \bar{X})^2 - [\sum (x + \bar{X})]^2} \\ &= \frac{N \sum (xy + x\bar{Y} + \bar{X}y + \bar{X}\bar{Y}) - (\sum x + N\bar{X})(\sum y + N\bar{Y})}{N \sum (x^2 + 2x\bar{X} + \bar{X}^2) - (\sum x + N\bar{X})^2} \end{aligned}$$

$$= \frac{N \sum xy + N\bar{Y} \sum x + N\bar{X} \sum y + N^2 \bar{X} \bar{Y} - (\sum x + N\bar{X})(\sum y + N\bar{Y})}{N \sum x^2 + 2N\bar{X} \sum x + N^2 \bar{X}^2 - (\sum x + N\bar{X})^2}$$

但 $\sum x = \sum (X - \bar{X}) = 0$, 并且 $\sum y = \sum (Y - \bar{Y}) = 0$, 因此上式可简写为

$$a_1 = \frac{N \sum xy + N^2 \bar{X} \bar{Y} - N^2 \bar{X} \bar{Y}}{N \sum x^2 + N^2 \bar{X}^2 - N^2 \bar{X}^2} = \frac{\sum xy}{\sum x^2}$$

或

$$\begin{aligned} a_1 &= \frac{\sum xy}{\sum x^2} = \frac{\sum x(Y - \bar{Y})}{\sum x^2} \\ &= \frac{\sum xY - \bar{Y} \sum x}{\sum x^2} = \frac{\sum xY}{\sum x^2} \end{aligned}$$

由(37)式知最小二乘直线为 $y = a_1 x$, 即

$$y = \left(\frac{\sum xy}{\sum x^2} \right) x \quad \text{或} \quad y = \left(\frac{\sum xY}{\sum x^2} \right) x$$

(b) 若 $\bar{X} = 0, x = X - \bar{X} = X$, 则由

$$y = \left(\frac{\sum xY}{\sum x^2} \right) x$$

得

$$y = \left(\frac{\sum XY}{\sum X^2} \right) X \quad \text{或} \quad Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2} \right) X$$

(c) 通过交换 X 与 Y 或 x 与 y , 则可得方程

$$x = \left(\frac{\sum xy}{\sum y^2} \right) y$$

(d) 由(a)知方程为

$$y = \left(\frac{\sum xy}{\sum x^2} \right) x \tag{39}$$

由(c)知方程为

$$x = \left(\frac{\sum xy}{\sum y^2} \right) y$$

或

$$y = \left(\frac{\sum y^2}{\sum xy} \right) x \tag{40}$$


因为一般

$$\frac{\sum xy}{\sum x^2} \neq \frac{\sum y^2}{\sum xy}$$

故直线(39)和(40)一般情况下是不同的, 然而, 它们相交于点 $x=0, y=0$ (即点 (\bar{X}, \bar{Y})).

13.16 若 $X' = X + A, Y' = Y + B$, 其中 A 与 B 是任意常数, 证明:

$$a_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} = \frac{N \sum X'Y' - (\sum X')(\sum Y')}{N \sum X'^2 - (\sum X')^2} = a'_1$$

证明 

$$x' = X' - \bar{X}' = (X + A) - (\bar{X} + A) = X - \bar{X} = x$$

$$y' = Y' - \bar{Y}' = (Y + B) - (\bar{Y} + B) = Y - \bar{Y} = y$$

$$\frac{\sum xy}{\sum x^2} = \frac{\sum x'y'}{\sum x'^2}$$

由习题 13.15, 即可得证结论, 类似的结果对 b_1 成立.

这个结论是有用的, 它可简化回归直线的计算(见习题 13.17 的解法二).

注意, 若 $X' = c_1 X + A$, $Y' = c_2 Y + B$, 结论将不成立, 除非 $c_1 = c_2$.

13.17 对习题 13.10 中数据拟合一最小二乘直线, 其中设(a) X 为自变量, (b) Y 为因变量.

解 **解法一** (a) 由习题 13.15(a) 结论, 知所求直线为

$$y = \left(\frac{\sum xy}{\sum x^2} \right) x$$

其中 $x = X - \bar{X}$, $y = Y - \bar{Y}$. 一些涉及和的计算值由表 13.6 给出, 则有

$$\bar{X} = 802/12 = 66.8, \quad \bar{Y} = 1850/12 = 154.2.$$

表 13.6

身高 X	体重 Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	xy	x^2	y^2
70	155	3.2	0.8	2.56	10.24	0.64
63	150	-3.8	-4.2	15.96	14.44	17.64
72	180	5.2	25.8	134.16	27.04	665.64
60	135	-6.8	-19.2	130.56	46.24	368.64
66	156	-0.8	1.8	-1.44	0.64	3.24
70	168	3.2	13.8	44.16	10.24	190.44
74	178	7.2	23.8	171.36	51.84	566.44
65	160	-1.8	5.8	-10.44	3.24	33.64
62	132	-4.8	-22.2	106.56	23.04	492.84
67	145	0.2	-9.2	-1.84	0.04	84.64
65	139	-1.8	-15.2	27.36	3.24	231.04
68	152	1.2	-2.2	-2.64	1.44	4.84
$\sum X = 802$ $\bar{X} = 66.8$	$\sum Y = 1850$ $\bar{Y} = 154.2$			$\sum xy = 616.32$	$\sum x^2 = 191.68$	$\sum y^2 = 2659.68$

则所求直线为

$$y = \frac{616.32}{191.68} x = 3.22x$$

或写成

$$Y - 154.2 = 3.22(X - 66.8) \quad \text{或} \quad Y = 3.22X - 60.9$$

该方程称为 Y 关于 X 的回归直线, 用于对给定 X 值估计 Y 值.

(b) 若 Y 是因变量, 则方程为

$$x = \left(\frac{\sum xy}{\sum y^2} \right) y = \frac{616.32}{2659.68} y = 0.232y$$

也可写成

$$X = 31.0 + 0.232Y.$$

该方程称为 X 关于 Y 的回归直线, 用于对给定 Y 值估计 X 值.

注意, 习题 13.11 中方法也可用.

解法二 用习题 13.16 的结论, 我们可以从 X 与 Y 中减去适当的常数. 我们选择从 X 中减去 65, 从 Y 中减去 150, 则结论可以写成如表 13.7 所示.

$$a_1 = \frac{N \sum X'Y' - (\sum X')(\sum Y')}{N \sum X'^2 - (\sum X')^2} = \frac{12 \times 708 - 22 \times 50}{12 \times 232 - 22^2} = 3.22$$

$$b_1 = \frac{N \sum X'Y' - (\sum Y')(\sum X')}{N \sum Y'^2 - (\sum Y')^2} = \frac{12 \times 708 - 50 \times 22}{12 \times 2868 - 50^2} = 0.232$$

因为 $\bar{X} = 65 + 22/12 = 66.8$, $\bar{Y} = 150 + 50/12 = 154.2$, 则回归方程为

$$Y - 154.2 = 3.22(X - 66.8) \quad \text{和} \quad X - 66.8 = 0.232(Y - 154.2),$$

即为 $Y = 3.22X - 60.9$ 和 $X = 0.232Y + 31.0$, 这和解法一中结论一致。

表 13.7

X'	Y'	X^2	$X'Y'$	Y^2
5	5	25	25	25
-2	0	4	0	0
7	30	49	210	900
-5	-15	25	75	225
1	6	1	6	36
5	18	25	90	324
9	28	81	252	784
0	10	0	0	100
-3	-18	9	54	324
2	-5	4	-10	25
0	-11	0	0	121
3	2	9	6	4
$\sum X' = 22$	$\sum Y' = 50$	$\sum X'^2 = 232$	$\sum X'Y' = 708$	$\sum Y'^2 = 2868$

13.18 (a)在同一坐标系里,就习题 13.17 中数据画两条直线.

(b)若已知身高为 63 英寸,求体重为多少?

(c)若已知体重为 168 磅,求身高为多少?

解 (a)两条直线画出如图 13-10 所示,两直线相交点(\bar{X} , \bar{Y})即(66.8, 154.2).

(b)由 Y 关于 X 的回归直线 $Y = 3.22X - 60.9$, 当 $X = 63$ 时,

$$Y = 3.22 \times 63 - 60.9 = 142.$$

(c)由 X 关于 Y 的回归直线 $X = 31.0 + 0.232Y$, 当 $Y = 168$ 时,

$$X = 31.0 + 0.232 \times 168 = 70.0.$$

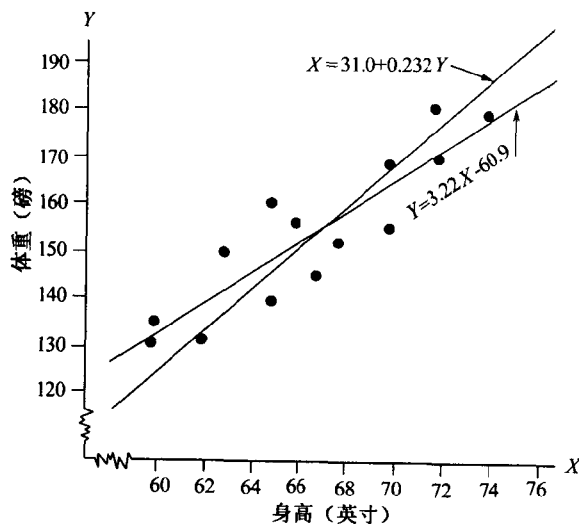


图 13-10

时间序列的应用

13.19 美国从 1989 年到 1995 年,农场实际土地总值(用 10 亿美元计算)如表 13.8 所示.用统计软件;

- (a)画出这些数据图；
 (b)求拟合这些数据的最小二乘直线方程；
 (c)估计美国 1988 年农场实际土地值,并和其实际价值 6268 亿美元做比较；
 (d)估计美国 1996 年农场土地价值,并和其实际值 8597 亿美元作比较.

表 13.8

年份	1989	1990	1991	1992	1993	1994	1995
总值	660.0	671.4	688.0	695.5	717.1	759.2	807.0

来源:美国农业部经济调查部.

解 (a)表 13.8 的数据图如图 3-11 中实线所示,图中虚线为最小二乘直线.

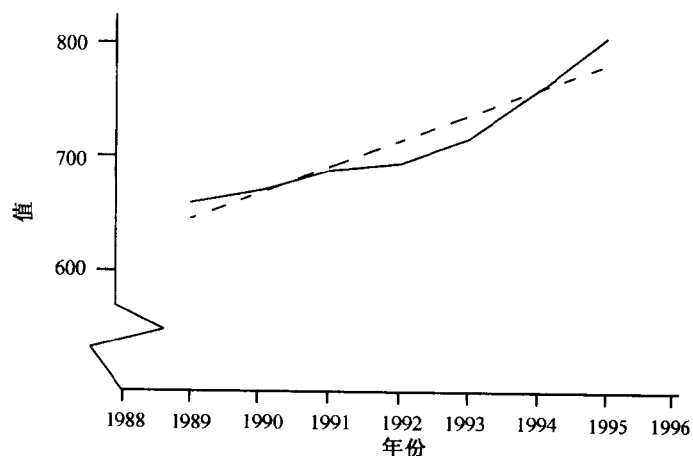


图 13-11 美国农场实际土地总值(10 亿美元)

(b)最小二乘直线的解由以下 Minitab 给出.

Data Display

Row	Year	Value
1	1989	660.0
2	1990	671.4
3	1991	688.0
4	1992	695.5
5	1993	717.1
6	1994	759.2
7	1995	807.0

MTB>regress c2 on 1 variable in c1

Regression Analysis

The regression equation is

$$\text{Value} = -45222.914286 + 23.060714 \text{ Year}$$

表 13.9 给出了表 13.8 中数据的拟合值和残差.把年份代入最小二乘直线的回归方程中可求出拟合值.例如, $-45222.914286 + 23.060714 \times 1989 = 644.846$.用实际的值减去拟合值即得残差.残差说明了最小二乘直线与实际数据的拟合情况.

表 13.9

年份	值	拟合值	残差
1989	660.0	644.846	15.1536
1990	671.4	667.907	3.4929
1991	688.0	690.968	-2.9679
1992	695.5	714.029	-18.5286
1993	717.1	737.089	-19.9893
1994	759.2	760.150	-0.9500
1995	807.0	783.211	23.7893

通常在数据分析前要对年份进行编码. 以下 Minitab 程序说明对年份用编码值的分析.

Data Display

Row	Year-coded	Value
1	0	660.0
2	1	671.4
3	2	688.0
4	3	695.5
5	4	717.1
6	5	759.2
7	6	807.0

MTB>regress c2 on 1 variable in c1

The regression equation is

Value = 644.846 + 23.061 Year-coded

表 13.10 给出了表 13.8 中数据对年份进行编码后的拟合值和残差.

表 13.10

年份编码	值	拟合值	残差
0	660.0	644.846	15.1536
1	671.4	667.907	3.4929
2	688.0	690.968	-2.9679
3	695.5	714.029	-18.5286
4	717.1	737.089	-19.9893
5	759.2	760.150	-0.9500
6	807.0	783.211	23.7893

(c)在(b)中得到的最小二乘方程可用来估计 1988 年农场实际土地总值. 对年份没有进行编码而得的方程计算得到 $-452229.14286 + 230.60714 \times 1988 = 6218$ 亿美元. 实际值为 6268 亿美元, 残差为 $6268 - 6218 = 5$ 亿美元. 对年份进行编码而得的方程计算得到 $6448.46 - 230.61 \times (-1) = 6218$ 亿美元. 注意在此编码中, 1988 记为 -1. (d)在(b)中得到的最小二乘方程可用来估计 1996 年农场实际土地总值. 对年份没有进行编码而得的方程计算得到 $-452229.14286 + 230.60714 \times 1996 = 8062.7$ 亿美元. 实际值为 8597 亿美元, 残差为 $8597 - 8062.7 = 534.3$ 亿美元. 对年份进行编码而得的方程计算得到 $6448.46 - 230.61 \times 7 = 6062.7$ 亿美元. 注意在此编码中, 1996 记为 7.

13.20 表 13.11 给出了由消费价格所测定的购买力情况.

(a)画出数据图.

(b)通过直接计算和用 Minitab 求趋势直线方程即拟合数据的最小二乘直线方程.

(c)假设趋势连续,预测 1998 年的购买力.

表 13.11

年份	1983	1984	1985	1986	1987	1988	1989
消费价格	1.003	0.961	0.928	0.913	0.880	0.846	0.807
年份	1990	1991	1992	1993	1994	1995	1996
消费价格	0.766	0.734	0.713	0.692	0.675	0.656	0.638

来源:美国劳工统计局,新近商业调查.

解 (a)如图 13-12 所示,实线为表 13.11 的数据图,虚线为最小二乘直线.

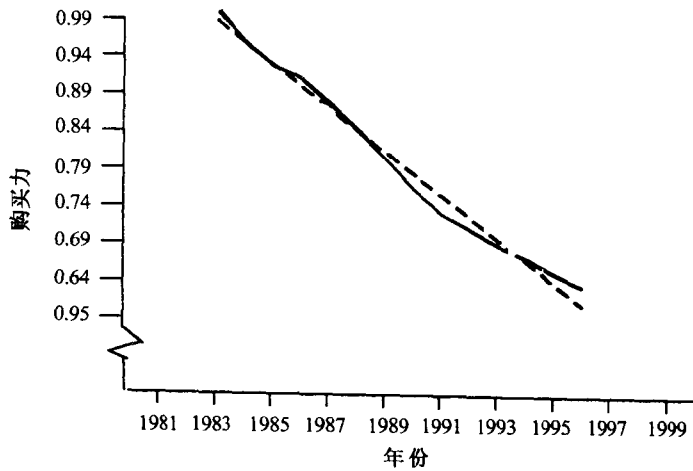


图 13-12

表 13.12

年份	X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy
1983	0	1.003	-6.5	0.202	42.25	-1.3130
1984	1	0.961	-5.5	0.160	30.25	-0.8800
1985	2	0.928	-4.5	0.127	20.25	-0.5715
1986	3	0.913	-3.5	0.112	12.25	-0.3920
1987	4	0.880	-2.5	0.079	6.25	-0.1975
1988	5	0.846	-1.5	0.045	2.25	-0.0675
1989	6	0.807	-0.5	0.006	0.25	-0.0030
1990	7	0.766	0.5	-0.035	0.25	-0.0175
1991	8	0.734	1.5	-0.067	2.25	-0.1005
1992	9	0.713	2.5	-0.088	6.25	-0.2200
1993	10	0.692	3.5	-0.109	12.25	-0.3815
1994	11	0.675	4.5	-0.126	20.25	-0.5670
1995	12	0.656	5.5	-0.145	30.25	-0.7975
1996	13	0.638	6.5	-0.163	42.25	-1.0595
	$\sum X = 91$ $\bar{X} = 6.5$	$\sum Y = 11.212$ $\bar{Y} = 0.801$			$\sum x^2 =$ 227.50	$\sum xy =$ -6.5680

(b)求趋势直线的计算如表 13.12 所示, 方程为 $y = \left(\frac{\sum xy}{\sum x^2} \right) x$, 其中 $x = X - \bar{X}$, $y = Y - \bar{Y}$, 则方程可写成

$$Y = -0.0289X + 0.9889$$

Minitab 解如下所示: 编码值 X 在 C1 列, 购买力值在 C2 列.

MTB>Regress 'Purchasing power' 1 'Year';

Regression Analysis

The regression equation is

Purchasing power = 0.989 - 0.0289 Year

(c)1998 年预计购买力为 $0.989 - 0.0289 \times 15 = 0.556$.

表 13.13

年份	购买价格	拟合值	残差
1983	1.003	0.989	0.014
1984	0.961	0.960	0.001
1985	0.928	0.931	-0.003
1986	0.913	0.902	0.011
1987	0.880	0.873	0.007
1988	0.846	0.844	0.002
1989	0.807	0.815	-0.008
1990	0.766	0.786	-0.020
1991	0.734	0.758	-0.024
1992	0.713	0.729	-0.016
1993	0.692	0.700	-0.008
1994	0.675	0.671	0.004
1995	0.656	0.642	0.014
1996	0.638	0.613	0.025

可简化为线性形式的非线性方程

13.21 表 13.14 给出了一给定气体随不同体积值 V 所相应的压强 P . 由热力学定理, P 与 V 之间有关系式: $PV^\gamma = C$, 其中 γ 与 C 是常数.

- 求 γ 和 C 的值;
- 写出关于 P 与 V 的方程;
- 当 $V = 100.0$ 立方英寸时, 估计 P 的值.

表 13.14

体积 V (立方英寸)	54.3	61.8	72.4	88.7	118.6	194.0
压强 P (磅/平方英寸)	61.2	49.2	37.6	28.4	19.2	10.1

解 由 $PV^\gamma = C$, 得

$$\log P + \gamma \log V = \log C \quad \text{或} \quad \log P = \log C - \gamma \log V$$

令 $\log V = X$, $\log P = Y$, 则上述方程式可写为

$$Y = a_0 + a_1 X \quad (41)$$

其中 $a_0 = \log C$, $a_1 = -\gamma$.

对应表 13.14 给出的 V 与 P 值, 表 13.15 给出了 $X = \log V$ 与 $Y = \log P$ 的值, 同时也给出了涉及最小二乘直线(41)式的计算值. 对应方程(41)的正规方程为

$$\sum Y = a_0 N + a_1 \sum X \quad \text{和} \quad \sum XY = a_0 \sum X + a_1 \sum X^2$$

可得

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2} = 4.20$$

$$a_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} = -1.40$$

因此 $Y = 4.20 - 1.40X$.

(a) 因为 $a_0 = 4.20 = \log C$, $a_1 = -1.40 = -\gamma$, 故 $C = 1.60 \times 10^4$, $\gamma = 1.40$.

(b) 所求 P, V 间方程为 $PV^{1.40} = 16000$.

(c) 当 $V = 100$ 时, $X = \log V = 2$, $Y = \log P = 4.20 - 1.40 \times 2 = 1.40$. 则

$$P = 10^{1.40} = 25.1 \text{ 磅/平方英寸}$$

表 13.15

$X = \log V$	$Y = \log P$	X^2	XY
1.7348	1.7868	3.0095	3.0997
1.7910	1.6946	3.2077	3.0350
1.8597	1.5752	3.4585	2.9294
1.9479	1.4533	3.7943	2.8309
2.0741	1.2833	4.3019	2.6617
2.2878	1.0043	5.2340	2.2976
$\sum X = 11.6953$	$\sum Y = 8.7975$	$\sum X^2 = 23.0059$	$\sum XY = 16.8543$

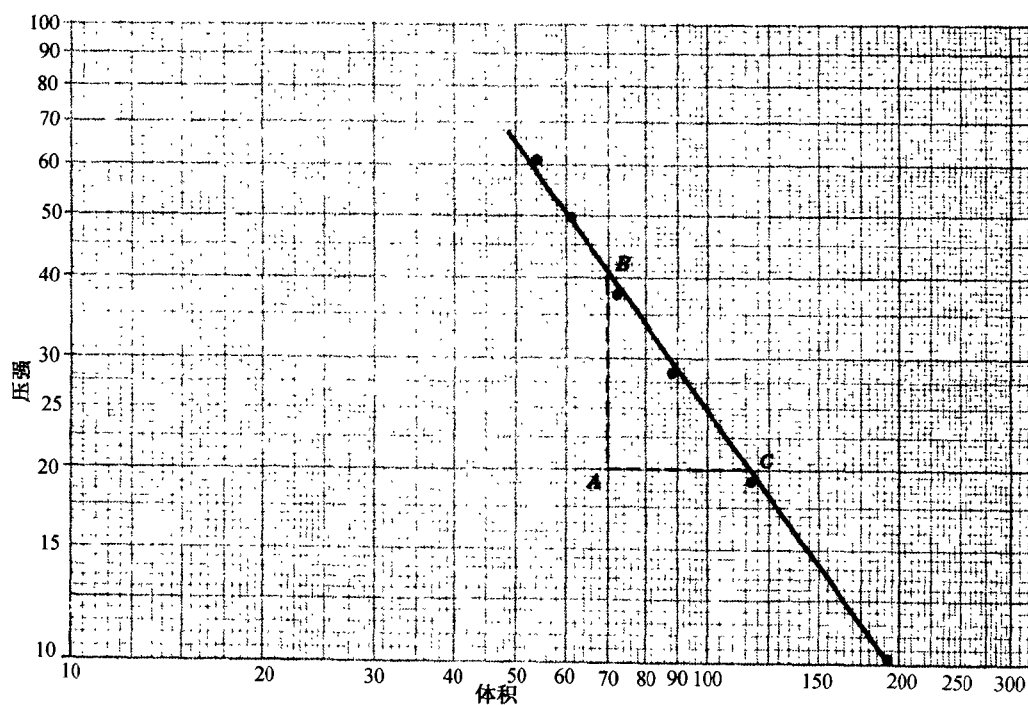


图 13-13

13.22 在对数-对数图解纸上描点解习题 13.21.

解 先由表 13.14 中每对值 (P, V) 求解相应的点 $(\log P, \log V)$. 再将这些点画在对数-对数图解纸上, 如图 13-13 所示. 然后画一直线接近这些点. 图中表明在 $\log P$ 与 $\log V$ 之间存在线性关系, 并可表示为

$$\log P = a_0 + a_1 \log V \quad \text{或} \quad Y = a_0 + a_1 X$$

斜率 a_1 是负的, 数值上等于 AB 与 AC 的长度比, 可得 $a_1 = -1.4$.

为得到 a_0 , 需要得到直线上的一点. 如当 $V = 100$ 时 $P = 25$, 因此, $a_0 = \log P - a_1 \log V = \log 25 + 1.4 \log 100 = 1.4 + 1.4 \times 2 = 4.2$, 故我们有 $\log P + 1.4 \log V = 4.2$, $\log PV^{1.4} = 4.2$, 即 $PV^{1.4} = 16000$.

最小二乘抛物线

13.23 表 13.16 给出了美国 1950 到 1995 年的总人口数, 以 5 年为一阶段, 单位以百万计算. 对这些数据拟合一直线, 同时拟合一抛物线, 对它们进行评论, 用两种模型来预测美国 2000 年的人口数.

表 13.16

年份	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995
人口	152	166	181	194	205	216	228	238	250	263

来源: 美国调查局.

解 由 Minitab 输出的最小二乘直线和最小二乘抛物线如下所示.

Row	Year	Population	x	xsquare
1	1950	152	0	0
2	1955	166	1	1
3	1960	181	2	4
4	1965	194	3	9
5	1970	205	4	16
6	1975	216	5	25
7	1980	228	6	36
8	1985	238	7	49
9	1990	250	8	64
10	1995	263	9	81

MTB>Regress 'population' on 1 predictor 'x'

Regression Analysis

The regression equation is

$$\text{Population} = 155 + 12.0 x$$

MTB>Regress 'population' on 2 predictors 'x' 'xsquare'

Regression Analysis

The regression equation is

$$\text{Population} = 153 + 13.6 x - 0.17.8 \text{xsquare}$$

表 13.17 给出了直线拟合值和残差.

表 13.17

年份	人口数	拟合值	残差
1950	152	155.164	-3.16364
1955	166	167.194	-1.19394
1960	181	179.224	1.77576
1965	194	191.255	2.74545

续表

年份	人口数	拟合值	残差
1970	205	203.285	1.71515
1975	216	215.315	0.68485
1980	228	227.345	0.65455
1985	238	239.376	-1.37576
1990	250	251.406	-1.40606
1995	263	263.436	-0.43636

表 13.18 给出了抛物线拟合的拟合值和残差.直线的残差平方和为 30.024,抛物线的残差平方和为13.289.显然,抛物线拟合数据比直线拟合得好.

表 13.18

年份	人口数	拟合值	残差
1950	152	153.027	-1.02727
1955	166	166.482	-0.48182
1960	181	179.580	1.41970
1965	194	192.323	1.67727
1970	205	204.709	0.29091
1975	216	216.739	-0.73939
1980	228	228.414	-0.41364
1985	238	239.732	-1.73182
1990	250	250.694	-0.69394
1995	263	261.300	1.70000

为了预测 2000 年人口数,要注意 2000 的编码值为 10.直线的预测值为 $155 + 12.0 \times 10 = 2.75$ 亿人,用抛物线模型预测的结果为 $153 + 13.6 \times 10 - 0.178 \times 100 = 2.712$ 亿人.

补充习题

直线

- 13.24 若 $3X + 2Y = 18$, 求(a)当 $Y = 3$ 时 X 的值, (b)当 $X = 2$ 时 Y 的值, (c)当 $Y = -5$ 时 X 的值, (d) $X = -1$ 时 Y 的值, (e) X 的截距, (f) Y 的截距.
- 13.25 在同一坐标系内, 画出方程(a) $Y = 3X - 5$ 和(b) $X + 2Y = 4$ 的图形, 问它们相交于哪一点?
- 13.26 (a)求通过点(3, -2)和(-1, 6)的直线方程;
(b)求(a)中直线的 X 截距和 Y 截距;
(c)分别求对应 $X = 3$ 和 $X = 5$ 的 Y 的值;
(d)直接从图形证明(a), (b), (c)的结论.
- 13.27 求斜率为 $\frac{2}{3}$, Y 截距为 -3 的直线方程.
- 13.28 (a)已知方程为 $3X - 5Y = 20$, 求其斜率和 Y 截距.
(b)求与(a)中直线平行且通过点(2, -1)的直线方程.
- 13.29 对通过点(5, 4)和(2, 8)的直线, 求(a)斜率, (b) Y 截距, (c)直线方程.
- 13.30 求 X 截距与 Y 截距分别为 3 和 -5 的直线方程.
- 13.31 已知 100 摄氏温度(100°C)相当于 212 华氏温度(212°F), 0°C 相当于 32°F , 假设摄氏温度与华氏温度间存在线性关系, 求(a)该直线方程, (b)对应 80°C 的华氏温度, (c)对应 68°F 的摄氏温度.

最小二乘直线

- 13.32 对表 13.19 中数据拟合一最小二乘直线, 其中设(a) X 为自变量, (b) X 为因变量. 在同一坐标系上画出数据点和直线.

表 13.19

X	3	5	6	8	9	11
Y	2	3	4	6	5	8

- 13.33 对习题 13.32 中数据,求(a)当 $X=5$ 和 $X=12$ 时相应的 Y 值,(b)当 $Y=7$ 时 X 的值.
- 13.34 (a)用徒手方法,求拟合习题 13.32 中数据的直线方程;(b)用(a)中的结论,回答习题 13.33 中问题.
- 13.35 表 13.20 给出了从一大群学生中随机抽取的 10 名学生的代数与物理的期末成绩.
- (a)画出数据图;
- (b)设 X 为自变量,求拟合这些数据的最小二乘直线;
- (c)设 Y 为自变量,求拟合这些数据的最小二乘直线;
- (d)若一学生代数成绩为 75,则他的物理期望分为多少?
- (e) 若一学生物理成绩为 95,则他代数的期望分为多少?

表 13.20

代数(X)	75	80	93	65	87	71	98	68	84	77
物理(Y)	82	78	86	72	91	80	95	72	89	74

- 13.36 表 13.21 给出在 1990~1996 年间每 1000 人的出生率.
- (a)画出数据图;
- (b)求拟合数据的最小二乘直线,把 1990 到 1996 年记为 0 到 6;
- (c)计算趋势值(拟合值)和残差;
- (d)假设目前趋势连续,预测 2000 年出生率.

表 13.21

年份	1990	1991	1992	1993	1994	1995	1996
人口出生率	16.6	16.3	15.9	15.5	15.2	14.8	14.5

来源:美国调查局.

- 13.37 表 13.22 给出美国 1985~1996 年的 85 岁及以上的人口数(单位以千记)
- (a) 画出数据图;
- (b)求拟合数据的最小二乘直线,把 1985 到 1996 记为数 0 到 11;
- (c)计算趋势值(拟合值)和残差;
- (d)假设目前趋势连续,预测 2005 年的 85 岁及以上的人口数.

表 13.22

年份	1985	1986	1987	1988	1989	1990
85 岁及以上	2667	2742	2823	2885	2968	3022
年份	1991	1992	1993	1994	1995	1996
85 岁及以上	3185	3306	3431	3541	3652	3762

来源:美国调查局.

- 13.38 对表 13.23 中数据,拟合一最小二乘抛物线 $Y = a_0 + a_1X + a_2X^2$.

表 13.23

X	0	1	2	3	4	5	6
Y	2.4	2.1	3.2	5.6	9.3	14.6	21.9

13.39 某人意识到危险后让汽车停下来的总时间等于反应时间(即从意识到危险到使用刹车这段时间)加上刹车时间(即刹车到车停住这段时间).表 13.24 给出了汽车从司机意识到危险的瞬时,其速度为 V ,到刹住车所跑的距离 D .

(a)画出 D 对应 V 的图形;

(b)对所给数据拟合一形式为 $D = a_0 + a_1 V + a_2 V^2$ 的最小二乘抛物线;

(c)求 $V = 45$ 米/小时和 $V = 80$ 米/小时时对应的 D 值.

表 13.24

速度 V (米/小时)	20	30	40	50	60	70
刹车距离 D (英尺)	54	90	138	206	292	396

13.40 表 13.25 给出了美国 1920~1980 年间男性人口与女性人口数,以 10 年为一阶段.

(a)画出口差异图;

(b)求拟合这些差异数据的最小二乘直线,记 1920 到 1990 为 0 到 7;

(c)假设趋势连续,估计 1995 年人口差距,然后与实际人口差距 5.75 作比较.问是否表明趋势是连续的?

表 13.25

年份	1920	1930	1940	1950	1960	1970	1980	1990
男性人口	53.90	62.14	66.06	75.19	88.33	98.93	110.05	121.24
女性人口	51.81	60.64	65.61	76.14	90.99	104.31	116.49	127.47
差	-2.09	-1.50	-0.45	0.95	2.66	5.38	6.44	6.23

来源:美国调查局.

13.41 用男性人口与女性人口的比率讨论习题 13.40.

13.42 对习题 13.40 中差异数据拟合一最小二乘抛物线.

13.43 表 13.26 给出了在 X 小时中,每单位容积所培育的细菌数 Y 的值.

表 13.26

小时数(X)	0	1	2	3	4	5	6
每单位容积的细菌数(Y)	32	47	65	92	132	190	275

(a)在半对数图解纸上画出数据图,对 Y 用对数尺度,对 X 用算术尺度;

(b)对数据拟合一形式为 $Y = ab^x$ 的最小二乘曲线,并解释这特殊方程为什么产生如此好的结果;

(c)把从方程中所得 Y 值与实际值作比较;

(d)当 $X = 7$ 时估计 Y 的值.

13.44 在习题 13.43 中,证明不用最小二乘法,怎样从一半对数图纸上的图得到所需方程.

第十四章 相关理论

相关与回归

在上一章中,我们讨论了从一个或多个相关变量(自变量)得到一个因变量的估计或回归问题.本章我们将进一步讨论变量间的相关程度,以确定一直线方程或其他方程是如何好地描述或表达变量间的关系.

若变量的所有值都完全满足一个方程,则说这些变量是**完全相关**的.圆的周长 C 与半径 r 有关系式 $c = 2\pi r$,因此说,周长 C 与半径 r 是完全相关的.若把两个骰子同时投掷 100 次,其每次投出的相应点之间没有任何关系(除非这些投掷是负重的),则我们说他们是**不相关**的.而像人的身高和体重这样的变量间则存在**某种**关系.

当只考虑两个变量时,属于**单相关**或**单回归**问题.当考虑两个以上变量时,属于**多重相关**或**多重回归**问题.本章我们只考虑单相关问题,多重相关或回归的问题将在第十五章讨论.

线性相关

若考虑两个变量 X 和 Y ,散点图给出了 (X, Y) 在直角坐标系中的位置.若散点图中的所有点看上去都在一条直线附近波动,如图 14-1(a)和 14-1(b)所示,则称变量间是**线性相关**的.此时,正如我们在第十三章中所看到的,可用一直线方程进行回归或估计.

若 Y 随着 X 的增加而增加,如图 14-1(a),则此相关称为**正相关**或**直接相关**;若 Y 随着 X 的增加而减少,如图 14-1(b),则此相关称为**负相关**或**逆相关**.

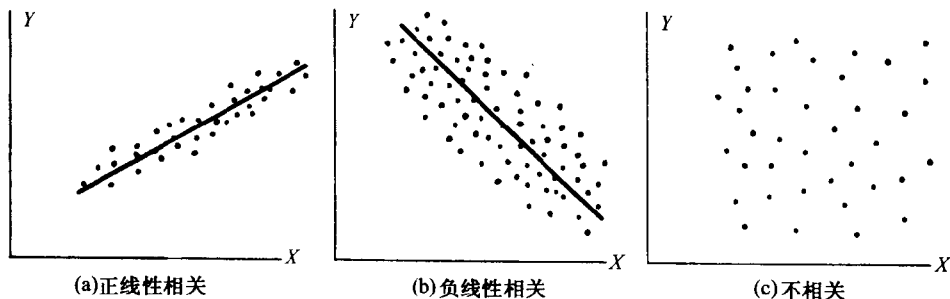


图 14-1

若所有的点看上去都在某条曲线的周围波动,则称此相关为**非线性相关**,此时可用一非线性方程进行回归讨论,如第十三章所示.很显然非线性相关也会有时是正相关,有时是负相关.

如图 14-1(c),若变量间没有显示出任何关系,则称变量间**不相关**.

相关性度量

用**定性**的方法,我们可以通过观察散点图来判定用一给定直线或曲线描述变量之间关系的好坏程度.例如,比起图 14-1(b)来,可以看出一条直线能更好地描述 14-1(a)中 X 和 Y 的关系,因为图 14-1(a)中的直线周围的点的分散程度不是很大.

如果想用**定量**的方法来讨论样本数据关于直线或曲线的散布问题,则有必要设计一种相关性度量方法.

最小二乘回归直线

我们首先考虑用直线描述两变量之间关系的好坏程度问题. 为此, 我们将用到第十三章中所得到的最小二乘回归直线方程. 正如我们所看到的, Y 关于 X 的最小二乘回归直线方程为

$$Y = a_0 + a_1 X \quad (1)$$

其中 a_0 和 a_1 由如下正规方程

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned} \quad (2)$$

计算, 因此

$$\begin{aligned} a_0 &= \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2} \\ a_1 &= \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} \end{aligned} \quad (3)$$

同理, X 关于 Y 的最小二乘回归直线方程为

$$X = b_0 + b_1 Y \quad (4)$$

其中 b_0 和 b_1 由如下正规方程

$$\begin{aligned} \sum X &= b_0 N + b_1 \sum Y \\ \sum XY &= b_0 \sum Y + b_1 \sum Y^2 \end{aligned} \quad (5)$$

计算, 因此

$$\begin{aligned} b_0 &= \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N \sum Y^2 - (\sum Y)^2} \\ b_1 &= \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum Y^2 - (\sum Y)^2} \end{aligned} \quad (6)$$

方程(1)和(4)也可分别用如下形式表示

$$y = \left[\frac{\sum xy}{\sum x^2} \right] x \quad \text{和} \quad x = \left[\frac{\sum xy}{\sum y^2} \right] y \quad (7)$$

其中 $x = X - \bar{X}$, $y = Y - \bar{Y}$.

当且仅当图中的点都在直线上时, 这两个回归方程才相同, 此时, 我们说 X 与 Y 之间是**完全线性相关**.

估计的标准误差

若把从方程(1)中由给定的 X 估计出的 Y 值记为 Y_{est} , 则 Y 关于 X 的回归直线的散布度量由

$$s_{Y.X} = \sqrt{\frac{\sum (Y - Y_{\text{est}})^2}{N}} \quad (8)$$

给出, 称为 Y 关于 X 的估计的标准误差.

若采用回归直线(4), 可同理给出 X 关于 Y 的估计的标准误差

$$s_{X.Y} = \sqrt{\frac{\sum (X - X_{\text{est}})^2}{N}} \quad (9)$$

一般说来, $s_{Y.X} \neq s_{X.Y}$.

为了便于计算,方程(8)一般写成下列形式(见习题 14.3):

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N} \quad (10)$$

方程(9)也有类似的表达式.

估计的标准误差与标准差的性质类似.例如,若在与回归直线相距 $s_{Y.X}$, $2s_{Y.X}$, $3s_{Y.X}$ 处分别作与其平行的直线,我们将发现,若 N 足够大,将有 68%, 95%, 99.7% 的样本点落在这些平行线之间.

修正的标准差

$$\hat{s} = \sqrt{\frac{N}{N-1}} \cdot s$$

对小样本数据是十分有用的,修正的估计的标准误差

$$\hat{s}_{Y.X} = \sqrt{\frac{N}{N-2}} \cdot s_{Y.X}$$

也有如此作用.因此,一些统计学家在定义方程(8)或(9)时一般用 $N-2$ 替换 N .

回归平方和与残差平方和

Y 的总变差定义为 $\sum (Y - \bar{Y})^2$, 即 Y 与均值 \bar{Y} 的偏差的平方和. 正如习题 14.7 中所示,总变差还有如下表述:

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y_{\text{est}})^2 + \sum (Y_{\text{est}} - \bar{Y})^2 \quad (11)$$

方程(11)右边的第一项称为残差平方和,第二项称为回归平方和.这里偏差 $Y_{\text{est}} - \bar{Y}$ 有确定的形式,而偏差 $Y - Y_{\text{est}}$ 则是随机的或不可预见的.对变量 X 也有相同的结论.

相关系数

回归平方和与总变差之比称为判定系数.若回归平方和为 0,则比值为 0;若残差平方和为 0,则比值为 1.也就是说比值只能在 0 与 1 之间取值.因为比值总是非负的,我们一般把它记为 r^2 .数 r 称为相关系数,由下式表示

$$r = \pm \sqrt{\frac{\text{回归平方和}}{\text{总变差}}} = \pm \sqrt{\frac{\sum (Y_{\text{est}} - \bar{Y})^2}{\sum (Y - \bar{Y})^2}} \quad (12)$$

r 取值于 -1 与 +1 之间. +, - 两符号分别用来表示正线性相关和负线性相关.注意 r 是一无量纲的量,没有单位可言.

由(8)式和(11)式,可知 Y 的标准差为

$$s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}} \quad (13)$$

我们发现若不考虑符号,(12)式可以写成如下形式

$$r = \sqrt{1 - \frac{s_{Y.X}^2}{s_Y^2}} \quad \text{或} \quad s_{Y.X} = s_Y \sqrt{1 - r^2} \quad (14)$$

把 X 和 Y 对换一下,也可得到类似的方程.

在线性相关情形中,不管 X 与 Y 哪个是自变量,数 r 都是一样的.因此说 r 是两变量间线性相关的非常好的度量.

关于相关系数的附注

(12)式和(14)式中定义的相关系数很具有一般性,既可用于线性关系,也可用于非线性关系,其惟一不同之处是 Y_{est} 是由一非线性回归方程计算求得而不是由一线性回归方程求得.另

外, 在非线性关系中, $+$, $-$ 符号也可忽略. 定义估计的标准误差的(8)式也具有一般性. 然而, (10)式只能用于线性回归. 若用于其他回归, 则必须修正一下. 例如, 若估计的方程为

$$Y = a_0 + a_1X + a_2X^2 + \cdots + a_{n-1}X^{n-1} \quad (15)$$

则(10)式应修正为

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY - \cdots - a_{n-1} \sum X^{n-1} Y}{N} \quad (16)$$

此时, 修正的估计标准误差为

$$\hat{s}_{Y.X} = \sqrt{\frac{N}{N-n}} \cdot s_{Y.X}$$

其中 $N-n$ 称为自由度.

必须强调一下, 在每一种情形中所计算的 r 值都度量了相对于所假设的方程类型的相关程度. 因此, 若假设有一直线方程, 而方程(12)或(14)计算的值接近于 0, 这意味着变量间几乎没有线性关系. 然而, 这并非说明它们之间不存在任何关系, 因为变量间可能还存在高度非线性关系. 换句话说, 相关系数度量的是假设方程和数据之间的拟合优度. 除非特别说明, 一般情况下, 相关系数即指线性相关系数.

还应该说明一下, 高的相关系数值(即接近于 1 或 -1)未必描述了变量间的一种直接依赖关系. 例如每年出版的书籍的数目与每年的暴风雨的次数之间可能存在一种高度相关. 这种例子有时被认为是伪相关.

线性相关系数的积-矩公式

若假定两变量间存在线性关系, 则(12)式可写成

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} \quad (17)$$

其中 $x = X - \bar{X}$, $y = Y - \bar{Y}$ (见习题 14.10). 此公式称为积-矩公式, 它直接给出了 r 的符号, 还清楚地表明 X 和 Y 的对称性.

若分别记

$$s_{XY} = \frac{\sum xy}{N}, \quad s_X = \sqrt{\frac{\sum x^2}{N}}, \quad s_Y = \sqrt{\frac{\sum y^2}{N}} \quad (18)$$

则 s_X , s_Y 分别表示变量 X 与 Y 的标准差, s_X^2 与 s_Y^2 是它们的方差. 新的量 s_{XY} 称为 X 与 Y 的协方差. 用(18)式中的记法, 公式(17)可写成

$$r = \frac{s_{XY}}{s_X \cdot s_Y} \quad (19)$$

注意, r 不仅与 X 和 Y 的单位无关, 也与其初始值无关.

快捷计算公式

公式(17)能写成以下等价形式

$$r = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{[N \sum X^2 - (\sum X)^2] \cdot [N \sum Y^2 - (\sum Y)^2]}} \quad (20)$$

一般常用它来计算 r (见习题 14.15 和 14.16).

对二元频数表或二元频数分布中的分组数据(见习题 14.17), 用前几章所述的编码方法计算是相当便利的. 此时, 公式(20)又可写成

$$r = \frac{N \sum f u_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N \sum f_X u_X^2 - (\sum f_X u_X)^2] \cdot [N \sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}} \quad (21)$$

(见习题 14.18). 为了计算中更方便地使用公式(21), 一般要用到**相关表**(见习题 14.19).

对分组数据, 公式(18)可写成

$$s_{XY} = c_X c_Y \left[\frac{\sum f u_X u_Y}{N} - \left(\frac{\sum f_X u_X}{N} \right) \left(\frac{\sum f_Y u_Y}{N} \right) \right] \quad (22)$$

$$s_X = c_X \sqrt{\frac{\sum f_X u_X^2}{N} - \left(\frac{\sum f_X u_X}{N} \right)^2} \quad (23)$$

$$s_Y = c_Y \sqrt{\frac{\sum f_Y u_Y^2}{N} - \left(\frac{\sum f_Y u_Y}{N} \right)^2} \quad (24)$$

其中 c_X, c_Y 分别表示变量 X 与 Y 的组距宽度(假设是常量).

注意, (23)和(24)式等价于第四章的公式(11).

利用结论(22)~(24)式可看出, 公式(19)等价于公式(21).

回归直线和线性相关系数

Y 关于 X 的最小二乘回归直线 $Y = a_0 + a_1 X$ 可写成

$$Y - \bar{Y} = \frac{rs_Y}{s_X} (X - \bar{X}) \quad \text{或} \quad y = \frac{rs_Y}{s_X} x \quad (25)$$

同理, X 关于 Y 的最小二乘回归直线 $X = b_0 + b_1 Y$ 可写成

$$X - \bar{X} = \frac{rs_X}{s_Y} (Y - \bar{Y}) \quad \text{或} \quad x = \frac{rs_X}{s_Y} y \quad (26)$$

当且仅当 $r = \pm 1$ 时, 方程(25)和(26)的斜率相等, 此时, 这两条直线为同一条直线, 即 X 与 Y 之间存在完全线性相关. 若 $r = 0$, 则两直线垂直, X 与 Y 之间没有线性关系. 因此线性相关系数度量了两直线的偏离程度.

注意, 若方程(25)和(26)分别写成 $Y = a_0 + a_1 X$, $X = b_0 + b_1 Y$, 则有 $a_1 b_1 = r^2$ (见习题 14.22).

时间序列相关

若变量 X 与 Y 都与时间相关, 则在 X 与 Y 之间有可能存在某种关系, 尽管这种关系未必是一种直接依赖关系, 也可能是某种“伪相关”. 通过考虑相对于不同时刻的 (X, Y) 值, 与通常一样利用上面公式(见习题 14.28), 即可得到相关系数.

也有可能某个时刻的 X 值与前一时刻的 X 值存在某种相关, 这种相关通常称为**自相关**.

属性相关

本章中所讨论的方法不能用于考虑原本是非数值的变量间的相关性, 如人的某些属性(例如头发颜色, 眼睛颜色等). 对属性的相关性的讨论见第十二章.

相关的抽样理论

两个变量 X 与 Y 的所有可能成对数据构成的总体, 称为**二元总体**, 一般情况下我们假定其服从**二元正态分布**. 现从中抽取 N 对数据. 我们考虑总体的相关系数, 记为 ρ , 其值由样本相关系数 r 估计. 关于各个 ρ 值的显著性检验和假设检验需要有 r 的抽样分布知识. 对 $\rho = 0$, 此分布对称, 因此要用到 t 分布的统计量; 对 $\rho \neq 0$, 分布具有偏性, 此时由 Fisher 研究的一种转换将产生一个近似正态分布的统计量. 下面的检验概括了所涉及的过程:

1. $\rho = 0$ 的假设检验. 此时要用到这样一个事实, 即统计量

$$t = \frac{r \sqrt{N-2}}{\sqrt{1-r^2}} \quad (27)$$

服从自由度为 $\nu = N - 2$ 的 t 分布(见习题 14.31 和 14.32).

2. $\rho = \rho_0 \neq 0$ 的假设检验.此时要用到这样一个事实,即统计量

$$Z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) = 1.1513 \log \left(\frac{1+r}{1-r} \right) \quad (28)$$

(其中 $e = 2.71828 \dots$) 渐近于正态分布, 均值和标准差分别为

$$\mu_Z = \frac{1}{2} \ln \left(\frac{1+\rho_0}{1-\rho_0} \right) = 1.1513 \log \left(\frac{1+\rho_0}{1-\rho_0} \right), \quad \sigma_Z = \frac{1}{\sqrt{N-3}} \quad (29)$$

3. 相关系数间差异的显著性.为了确定样本容量分别为 N_1 和 N_2 的两个样本的相关系数 r_1 和 r_2 之间是否存在显著性差异, 我们必须用(28)式来分别计算相应于 r_1 和 r_2 的 Z_1 和 Z_2 值. 此时用到这样一个事实, 即检验统计量

$$z = \frac{Z_1 - Z_2 - \mu_{Z_1 - Z_2}}{\sigma_{Z_1 - Z_2}} \quad (30)$$

服从正态分布(见习题 14.35), 其中

$$\mu_{Z_1 - Z_2} = \mu_{Z_1} - \mu_{Z_2}$$

且

$$\sigma_{Z_1 - Z_2} = \sqrt{\sigma_{Z_1}^2 + \sigma_{Z_2}^2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

回归的抽样理论

回归方程 $Y = a_0 + a_1 X$ 是基于样本数据得到的. 我们通常对相应的总体的回归方程感兴趣. 下面给出了关于总体的三个检验:

1. $a_1 = A_1$ 的假设检验.为检验假设: 回归系数 a_1 等于某个特定值 A_1 , 需要用到这样一个事实, 即统计量

$$t = \frac{a_1 - A_1}{s_{Y \cdot X} / s_X} \sqrt{N - 2} \quad (31)$$

服从自由度为 $N - 2$ 的 t 分布. 这也可用来由样本值求总体回归系数的置信区间(见习题 14.36, 14.37).

2. 对预测值的假设检验.设 Y_0 为当 X 取值 X_0 时, 利用样本回归方程计算出的 Y 值(即 $Y_0 = a_0 + a_1 X_0$), Y_P 为当 X 取值 X_0 时总体 Y 的预测值, 则统计量

$$t = \frac{Y_0 - Y_P}{s_{Y \cdot X} \sqrt{N + 1 + (X_0 - \bar{X})^2 / s_X^2}} \sqrt{N - 2} = \frac{Y_0 - Y_P}{\hat{s}_{Y \cdot X} \sqrt{1 + 1/N + (X_0 - \bar{X})^2 / (N s_X^2)}} \quad (32)$$

服从自由度为 $N - 2$ 的 t 分布. 由此也能求出总体预测值的置信限(见习题 14.38).

3. 对预测均值的假设检验.设 Y_0 为当 X 取值 X_0 时, 利用样本回归方程计算出的 Y 值(即 $Y_0 = a_0 + a_1 X_0$), \bar{Y}_P 为当 X 取值 X_0 时总体 Y 的预测均值, 则统计量

$$t = \frac{Y_0 - \bar{Y}_P}{s_{Y \cdot X} \sqrt{1 + (X_0 - \bar{X})^2 / s_X^2}} \sqrt{N - 2} = \frac{Y_0 - \bar{Y}_P}{\hat{s}_{Y \cdot X} \sqrt{1/N + (X_0 - \bar{X})^2 / (N s_X^2)}} \quad (33)$$

服从自由度为 $N - 2$ 的 t 分布. 由此也能求出总体预测均值的置信限(见习题 14.39).

习题及解答

散点图和回归直线

14.1 表 14.1 给出了 12 对父与子(大儿子)身高 X 与 Y 的样本值, 单位为英寸.

- (a) 构造散点图;
 (b) 求 Y 关于 X 的最小二乘回归直线;
 (c) 求 X 关于 Y 的最小二乘回归直线.

表 14.1

父亲身高 X (英寸)	65	63	67	64	68	62	70	66	68	67	69	71
儿子身高 Y (英寸)	68	66	68	65	69	66	68	65	71	67	68	70

解 (a) 把点 (X, Y) 描在直角坐标系内即得散点图, 如图 14-2 所示.

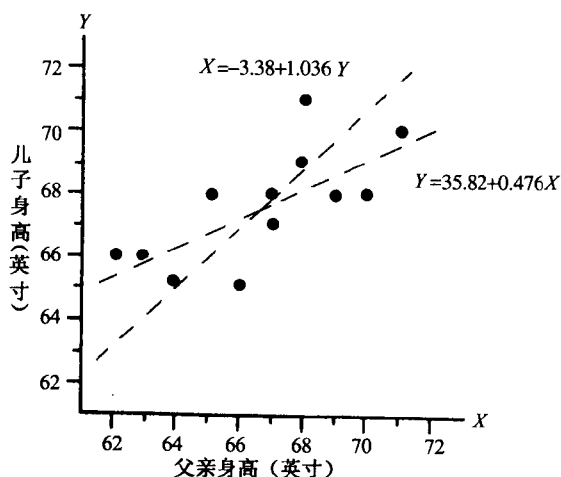


图 14-2

(b) Y 关于 X 的回归直线为 $Y = a_0 + a_1 X$, 其中 a_0 和 a_1 通过解如下正规方程

$$\sum Y = a_0 N + a_1 \sum X$$

$$\sum XY = a_0 \sum X + a_1 \sum X^2$$

可得. 求和的值由表 14.2 给出. 因此上述方程变为

$$12a_0 + 800a_1 = 811$$

$$800a_0 + 53418a_1 = 54107$$

通过解方程可知 $a_0 = 35.82$, $a_1 = 0.476$. 故 $Y = 35.82 + 0.476X$. 该直线见图 14-2.

另解

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2} = 35.82,$$

$$a_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} = 0.476$$

(c) X 关于 Y 的回归方程为 $X = b_0 + b_1 Y$, 其中 b_0 和 b_1 通过解如下正规方程

$$\sum X = b_0 N + b_1 \sum Y$$

$$\sum XY = b_0 \sum Y + b_1 \sum Y^2$$

得到. 代入表 14.2 中的数据, 上述方程变为

$$12b_0 + 811b_1 = 800$$

$$811b_0 + 54849b_1 = 54107$$

从而解得 $b_0 = -3.38$, $b_1 = 1.036$, 故 $X = -3.38 + 1.036Y$. 直线图见图 14-2.

另解

$$b_0 = \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N \sum Y^2 - (\sum Y)^2} = -3.38,$$

$$b_1 = \frac{N \sum XY - (\sum Y)(\sum X)}{N \sum Y^2 - (\sum Y)^2} = 1.036$$

表 14.2

X	Y	X ²	XY	Y ²
65	68	4225	4420	4624
63	66	3969	4158	4356
67	68	4489	4556	4624
64	65	4096	4160	4225
68	69	4624	4692	4761
62	66	3844	4092	4356
70	68	4900	4760	4624
66	65	4356	4290	4225
68	71	4624	4828	5041
67	67	4489	4489	4489
69	68	4761	4692	4624
71	70	5041	4970	4900
$\sum X = 800$	$\sum Y = 811$	$\sum X^2 = 53418$	$\sum XY = 54107$	$\sum Y^2 = 54849$

14.2 用 Minitab 求解习题 14.1. 构造拟合值 Y_{est} 和残差的数据表. 分别对两条回归直线求残差平方和.

解 表 14.3 给出了 Y 关于 X 的回归直线的拟合值, 残差和残差平方.

表 14.3

X	Y	拟合值 Y_{est}	残差 $Y - Y_{\text{est}}$	残差平方
65	68	66.79	1.21	1.47
63	66	65.84	0.16	0.03
67	68	67.74	0.26	0.07
64	65	66.31	-1.31	1.72
68	69	68.22	0.78	0.61
62	66	65.36	0.64	0.41
70	68	69.17	-1.17	1.37
66	65	67.27	-2.27	5.13
68	71	68.22	2.78	7.74
67	67	67.74	-0.74	0.55
69	68	68.89	-0.69	0.48
71	70	69.95	0.35	0.12
			和 = 0	和 = 19.70

Y 关于 X 的最小二乘回归直线的 Minitab 输出如下:
MTB>Regress 'Y' on 1 predictor 'X'
Regression Analysis

The regression equation is $Y = 35.8 + 0.476X$

X 关于 Y 的最小二乘回归直线的 Minitab 输出如下:

MTB>Regress 'X' on 1 predictor 'Y'

Regression Analysis

The regression equation is $X = -3.4 + 1.04Y$

表 14.4 给出了 X 关于 Y 的回归直线的拟合值, 残差和残差平方.

表 14.4

X	Y	拟合值 X_{est}	残差 $X - X_{\text{est}}$	残差平方
65	68	67.10	-2.10	4.40
63	66	65.03	-2.03	4.10
67	68	67.10	-0.10	0.01
64	65	63.99	0.01	0.00
68	69	68.13	-0.13	0.02
62	66	65.03	-3.03	9.15
70	68	67.10	2.90	8.42
66	65	63.99	2.01	4.04
68	71	70.21	-2.21	4.87
67	67	66.06	0.94	0.88
69	68	67.10	1.90	3.62
71	70	69.17	1.83	3.34
			和 = 0	和 = 42.85

通过比较残差平方和可知, Y 关于 X 的最小二乘回归线的拟合值要比 X 关于 Y 的最小二乘回归线的拟合值好得多. 前面曾讲到过, 残差平方和越小, 回归模型和数据的拟合就越好. 因此可知, 由父亲身高来估计儿子身高显然比由儿子身高估计父亲身高更合适.

估计的标准误差

14.3 若 Y 关于 X 的回归直线为 $Y = a_0 + a_1X$, 证明估计的标准误差 $s_{Y.X}$ 由下式给出.

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N}$$

解 从回归直线计算的 Y 的估计值为 $Y_{\text{est}} = a_0 + a_1X$, 故

$$\begin{aligned} s_{Y.X}^2 &= \frac{\sum (Y - Y_{\text{est}})^2}{N} = \frac{\sum (Y - a_0 - a_1X)^2}{N} \\ &= \frac{\sum Y(Y - a_0 - a_1X) - a_0 \sum (Y - a_0 - a_1X) - a_1 \sum X(Y - a_0 - a_1X)}{N} \end{aligned}$$

但

$$\sum (Y - a_0 - a_1X) = \sum Y - a_0N - a_1 \sum X = 0$$

且

$$\sum X(Y - a_0 - a_1X) = \sum XY - a_0 \sum X - a_1 \sum X^2 = 0$$

它们是从如下正规方程推导而来

$$\begin{aligned} \sum Y &= a_0N + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned}$$

因此

$$s_{Y.X}^2 = \frac{\sum Y(Y - a_0 - a_1X)}{N} = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N}$$

该结果可以推广到非线性回归方程.

14.4 若 $x = X - \bar{X}$, $y = Y - \bar{Y}$, 证明习题 14.3 中结论可写成

$$s_{Y.X}^2 = \frac{\sum y^2 - a_1 \sum xy}{N}$$

解 由 $X = x + \bar{X}$, $Y = y + \bar{Y}$, 利用习题 14.3 中结论, 有

$$\begin{aligned} Ns_{Y.X}^2 &= \sum Y^2 - a_0 \sum Y - a_1 \sum XY \\ &= \sum (y + \bar{Y})^2 - a_0 \sum (y + \bar{Y}) - a_1 \sum (x + \bar{X})(y + \bar{Y}) \\ &= \sum (y^2 + 2y\bar{Y} + \bar{Y}^2) - a_0(\sum Y + N\bar{Y}) \\ &\quad - a_1 \sum (xy + \bar{X}y + x\bar{Y} + \bar{X}\bar{Y}) \\ &= \sum y^2 + 2\bar{Y} \sum Y + N\bar{Y}^2 - a_0 N\bar{Y} - a_1 \sum xy - a_1 \bar{X} \sum y \\ &\quad - a_1 \bar{Y} \sum x - a_1 N\bar{X}\bar{Y} \\ &= \sum y^2 + N\bar{Y}^2 - a_0 N\bar{Y} - a_1 \sum xy - a_1 N\bar{X}\bar{Y} \\ &= \sum y^2 - a_1 \sum xy + N\bar{Y}(\bar{Y} - a_0 - a_1 \bar{X}) \\ &= \sum y^2 - a_1 \sum xy \end{aligned}$$

其中我们用到结论 $\sum x = 0$, $\sum y = 0$, 及 $\bar{Y} = a_0 + a_1 \bar{X}$ (由正规方程 $\sum Y = a_0 N + a_1 \sum X$ 两边同除以 N 而得).

14.5 对习题 14.1 中的数据计算标准误差 $s_{Y.X}$. 分别用 (a) 定义, (b) 习题 14.4 的结论.

解 (a) 从习题 14.4(b) 知, Y 关于 X 的回归直线为 $Y = 35.82 + 0.476X$. 表 14.5 给出了 Y 的真实值和估计值 (记为 Y_{est}), 该估计值由回归直线得到. 例如, 对应 $X = 65$, 有 $Y_{\text{est}} = 35.82 + 0.476 \times 65 = 66.76$. 表 14.5 同时列出了 $Y - Y_{\text{est}}$ 的值, 以便计算 $s_{Y.X}$.

$$s_{Y.X}^2 = \frac{\sum (Y - Y_{\text{est}})^2}{N} = \frac{1.24^2 + 0.19^2 + \cdots + 0.38^2}{12} = 1.642$$

故 $s_{Y.X} = \sqrt{1.643} = 1.28$ 英寸.

(b) 从习题 14.1, 14.2 和 14.4 有

$$s_{Y.X}^2 = \frac{\sum y^2 - a_1 \sum xy}{N} = \frac{38.92 - 0.476 \times 40.34}{12} = 1.643$$

故 $s_{Y.X} = \sqrt{1.643} = 1.28$ 英寸.

表 14.5

X	65	63	67	64	68	62	70	66	68	67	69	71
Y	68	66	68	65	69	66	68	65	71	67	68	70
Y_{est}	66.76	65.81	67.71	66.28	68.19	65.33	69.14	67.24	68.19	67.71	68.66	69.62
$Y - Y_{\text{est}}$	1.24	0.19	0.29	-1.28	0.81	0.67	-1.14	-2.24	2.81	-0.71	-0.66	0.38

14.6 (a) 构造两条与习题 14.1 中回归直线平行, 且距离为 $s_{Y.X}$ 的直线;
(b) 确定位于两直线内的数据的百分数.

解 (a) 习题 14.1 中所得的回归直线为 $Y = 35.82 + 0.476X$, 在图 14-3 中用粗线表示, 两条与之平行且距离为 $s_{Y.X} = 1.28$ 的直线 (见习题 14.5), 在图 14-3 中用虚线表示.

(b) 从图 14-3 可看出, 12 个点中有 7 个点位于两直线之间, 3 个点好像在两直线上, 经进一步检验, 可知这 3 个点中有 2 个位于两直线之间, 故所求百分数为 $9/12 = 75\%$.

另解 从表 14.5 最后一行可知, 有 9 个点 (X, Y) 所对应的 $Y - Y_{\text{est}}$ 的值处于 -1.28 与 1.28 (即 $\pm s_{Y.X}$) 之间, 故所求的百分数为 $9/12 = 75\%$.

若这些点关于回归直线是正态分布的, 则理论上将应有 68% 的点处于两直线之间, 这一结论对大样本更适合.

注: 对习题 14.1 中身高的估计的标准误差的最佳估计可由

$$\hat{s}_{Y.X} = \sqrt{N/(N-2)} \cdot s_{Y.X} = \sqrt{12/10} \cdot 1.28 = 1.40 \text{ 英寸}$$

给出.

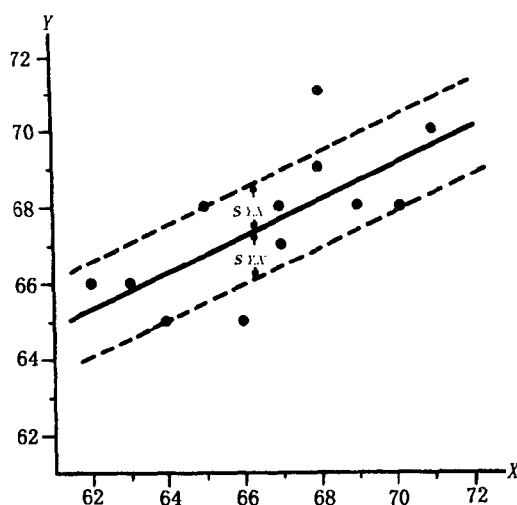


图 14-3

回归平方和与残差平方和

14.7 证明: $\sum (Y - \bar{Y})^2 = \sum (Y - Y_{\text{est}})^2 + \sum (Y_{\text{est}} - \bar{Y})^2$

证明 对 $Y - \bar{Y} = (Y - Y_{\text{est}}) + (Y_{\text{est}} - \bar{Y})$ 两边平方再求和, 则得

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y_{\text{est}})^2 + \sum (Y_{\text{est}} - \bar{Y})^2 + 2 \sum (Y - Y_{\text{est}})(Y_{\text{est}} - \bar{Y})$$

若能证明上式最后一项为 0, 则可得证结果. 因为

$$\begin{aligned} \sum (Y - Y_{\text{est}})(Y_{\text{est}} - \bar{Y}) &= \sum (Y - a_0 - a_1 X)(a_0 + a_1 X - \bar{Y}) \\ &= a_0 \sum (Y - a_0 - a_1 X) + a_1 \sum X(Y - a_0 - a_1 X) \\ &\quad - \bar{Y} \sum (Y - a_0 - a_1 X) \end{aligned}$$

而由正规方程知

$$\sum (Y - a_0 - a_1 X) = 0, \quad \sum X(Y - a_0 - a_1 X) = 0$$

因而得证.

用最小二乘曲线 $Y_{\text{est}} = a_0 + a_1 X + a_2 X^2 + \cdots + a_n X^n$, 可证以上结论对非线性回归同样成立.

14.8 对习题 14.1 中数据, 计算(a) 总变差, (b) 残差平方和, (c) 回归平方和.

解 最小二乘回归直线为 $Y_{\text{est}} = 35.8 + 0.476X$. 从表 14.6 可以看出, 总变差为 $\sum (Y - \bar{Y})^2 = 38.917$, 残差平方和 $\sum (Y - Y_{\text{est}})^2 = 19.703$, 回归平方和为 $\sum (Y_{\text{est}} - \bar{Y})^2 = 19.214$.

表 14.6

Y	Y_{est}	$(Y - \bar{Y})^2$	$(Y - Y_{\text{est}})^2$	$(Y_{\text{est}} - \bar{Y})^2$
68	66.7894	0.1739	1.46562	0.62985
66	65.8366	2.5059	0.02669	3.04986
68	67.7421	0.1739	0.06650	0.02532
65	66.3130	6.6719	1.72395	1.61292
69	68.2185	2.0079	0.61074	0.40387
66	65.3602	2.5059	0.40930	4.94068
68	69.1713	0.1739	1.37185	2.52257
65	67.2657	6.6719	5.13361	0.10065
71	68.2185	11.6759	7.73672	0.40387
67	67.7421	0.3399	0.55075	0.02532
68	68.6949	0.1739	0.48286	1.23628
70	69.6476	5.8419	0.12416	4.26273
$\bar{Y} = 67.5833$		和 = 38.917	和 = 19.703	和 = 19.214

下面的 Minitab 输出得到的结果与上同,用黑体字表示.注意,用软件使计算量减少了许多.

MTB>Regress 'Y' 1 'X'

SUBC>Constant;

SUBC>Brief 1.

Regression Analysis

The regression equation is

$$Y = 35.8 + 0.476X$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	19.214	19.214	9.75	0.011
Residual Error	10	19.703	1.970		
Total	11	38.917			

相关系数

14.9 用习题 14.8 中结论,求(a) 判定系数,(b) 相关系数.

解 (a) 判定系数 $= r^2 = \frac{\text{回归平方和}}{\text{总变差}} = \frac{19.214}{38.917} = 0.4937$

(b) 相关系数 $= r = \pm \sqrt{0.4937} = \pm 0.7027$

因为 X 与 Y 是直接相关的,所以选正号,且保留两个小数位,得 $r = 0.70$.

14.10 证明对线性回归而言,变量 X 与 Y 间的相关系数可写成

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

其中 $x = X - \bar{X}$, $y = Y - \bar{Y}$.

解 Y 关于 X 的最小二乘回归直线可写成 $Y_{\text{est}} = a_0 + a_1 X$ 或 $y_{\text{est}} = a_1 x$, 其中

$$a_1 = \frac{\sum xy}{\sum x^2}, \quad y_{\text{est}} = Y_{\text{est}} - \bar{Y}$$

则

$$\begin{aligned} r^2 &= \frac{\text{回归平方和}}{\text{总变差}} = \frac{\sum (Y_{\text{est}} - \bar{Y})^2}{\sum (Y - \bar{Y})^2} = \frac{\sum y_{\text{est}}^2}{\sum y^2} \\ &= \frac{\sum a_1^2 x^2}{\sum y^2} = \frac{a_1^2 \sum x^2}{\sum y^2} = \left(\frac{\sum xy}{\sum x^2} \right)^2 \frac{\sum x^2}{\sum y^2} = \frac{(\sum xy)^2}{(\sum x^2)(\sum y^2)} \end{aligned}$$

且

$$r = \pm \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

因为当 y_{est} 随 x 的增加而增加时, $\frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$ 是正的, 而当 y_{est} 随 x 的增加而减少时,

$\frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$ 是负的, 即式 $\frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$ 根据 y 与 x 的相关关系自动取号. 因此我们定义线性相关系数为

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

这通常称为线性相关系数的积-矩公式.

线性相关系数的积-矩公式

14.11 求表 14.7 中变量 X 和 Y 的线性相关系数.

表 14.7

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

解 计算的结果见表 14.8. 则

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{84}{\sqrt{132 \times 56}} = 0.977$$

正如我们在习题 13.8 和 13.12 中观察到的一样, r 的值表明 X 与 Y 间存在很高的相关性.

表 14.8

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy	y^2
1	1	-6	-4	36	24	16
3	2	-4	-3	16	12	9
4	4	-3	-1	9	3	1
6	4	-1	-1	1	1	1
8	5	1	0	1	0	0
9	7	2	2	4	4	4
11	8	4	3	16	12	9
14	9	7	4	49	28	16
$\sum X = 56$ $\bar{X} = 56/8 = 7$	$\sum Y = 40$ $\bar{Y} = 40/8 = 5$			$\sum x^2 = 132$	$\sum xy = 84$	$\sum y^2 = 56$

- 14.12 对习题 14.11 中的数据, 求(a) X 的标准差, (b) Y 的标准差, (c) X 的方差, (d) Y 的方差, (e) X 与 Y 的协方差. 并把这些值与 Minitab 结果进行比较, 解释其中存在的差异.

解 (a) X 的标准差为 $s_X = \sqrt{\frac{\sum (X - \bar{X})^2}{N}} = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{132}{8}} = 4.06$

(b) Y 的标准差为 $s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}} = \sqrt{\frac{\sum y^2}{N}} = \sqrt{\frac{56}{8}} = 2.65$

(c) X 的方差为 $s_X^2 = 16.50$

(d) Y 的方差为 $s_Y^2 = 7.00$

(e) X 与 Y 的协方差为 $s_{XY} = \frac{\sum xy}{N} = \frac{84}{8} = 10.5$

Minitab 输出如下:

MTB> Standard deviation c1

Column Standard deviation

Standard deviation of X = 4.3425

MTB> Standard deviation c2

Column Standard deviation

Standard deviation of Y = 2.8284

MTB> covariance c1 c2

Covariance

X

Y

X	18.85714	
Y	12.00000	8.00000

值之间的差别,主要是在 Minitab 中除以 $N-1$ 而非 N .

14.13 对习题 14.11 中的数据,验证公式 $r = \frac{s_{XY}}{s_X s_Y}$.

解 由习题 14.12 可知

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{10.50}{4.06 \times 2.65} = 0.976$$

除了舍入误差,所得结果与习题 14.11 中的结论一致.

14.14 用积-矩公式,求习题 14.1 中数据的线性相关系数.

解 由表 14.3 可知

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{40.34}{\sqrt{84.68 \times 38.92}} = 0.7027$$

与习题 14.9 中所得结论一致.

14.15 证明线性相关系数为

$$r = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{[N \sum X^2 - (\sum X)^2] \cdot [N \sum Y^2 - (\sum Y)^2]}}$$

证明 令 $x = X - \bar{X}$, $y = Y - \bar{Y}$, 由习题 14.10 中的结论,有

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{[\sum (X - \bar{X})^2] \cdot [\sum (Y - \bar{Y})^2]}} \quad (34)$$

因为 $\bar{X} = (\sum X)/N$, $\bar{Y} = (\sum Y)/N$, 所以

$$\begin{aligned} \sum (X - \bar{X})(Y - \bar{Y}) &= \sum (XY - \bar{X}Y - X\bar{Y} + \bar{X}\bar{Y}) \\ &= \sum XY - \bar{X} \sum Y - \bar{Y} \sum X + N\bar{X}\bar{Y} \\ &= \sum XY - N\bar{X}\bar{Y} - N\bar{Y}\bar{X} + N\bar{X}\bar{Y} \\ &= \sum XY + N\bar{X}\bar{Y} \\ &= \sum XY - \frac{(\sum X)(\sum Y)}{N} \end{aligned}$$

同理

$$\begin{aligned} \sum (X - \bar{X})^2 &= \sum (X^2 - 2X\bar{X} + \bar{X}^2) = \sum X^2 - 2\bar{X} \sum X + N\bar{X}^2 \\ &= \sum X^2 - \frac{2(\sum X)^2}{N} + \frac{(\sum X)^2}{N} \\ &= \sum X^2 - \frac{(\sum X)^2}{N} \end{aligned}$$

且

$$\sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{N}$$

则(34)可写成

$$\begin{aligned} r &= \frac{\sum XY - (\sum X)(\sum Y)/N}{\sqrt{[\sum X^2 - (\sum X)^2/N] \cdot [\sum Y^2 - (\sum Y)^2/N]}} \\ &= \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{[N \sum X^2 - (\sum X)^2] \cdot [N \sum Y^2 - (\sum Y)^2]}} \end{aligned}$$

14.16 用习题 14.15 中公式求习题 14.1 中数据的线性相关系数.

解

$$r = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{[N \sum X^2 - (\sum X)^2] \cdot [N \sum Y^2 - (\sum Y)^2]}}$$

$$= \frac{12 \times 54107 - 800 \times 811}{\sqrt{(12 \times 53418 - 800^2)(12 \times 54849 - 811^2)}} = 0.7027$$

结论与习题 14.9 和 14.14 的一致.

另解 r 的值和 X, Y 的初始值无关. 因此我们先进行数据变换: $X' = X - 60, Y' = Y - 60$, 然后按习题 14.15 中公式计算, 得

$$r = \frac{N \sum X'Y' - (\sum X')(\sum Y')}{\sqrt{[N \sum X'^2 - (\sum X')^2] \cdot [N \sum Y'^2 - (\sum Y')^2]}}$$

$$= \frac{12 \times 647 - 80 \times 91}{\sqrt{(12 \times 618 - 80^2)(12 \times 729 - 91^2)}} = 0.7027$$

分组数据的相关系数

14.17 表 14.9 给出了 100 名学生的数学、物理成绩的频数分布. 通过此表计算:

- 数学成绩在 70~79 之间且物理成绩在 80~89 之间的学生人数.
- 数学成绩低于 70 分的学生百分比.
- 物理成绩在 70 分以上而数学成绩在 80 分以下的学生人数.
- 假定及格线为 60 分, 则求至少有一门课及格的学生的百分比.

解 (a) 由表 14.9, 可得所求学生人数为 4.

(b) 数学成绩低于 70 分的学生人数 = (40~49 分的人数) + (50~59 分的人数) + (60~69 分的人数) = 7 + 15 + 25 = 47. 故百分比为 47/100 = 47%.

(c) 由表 14.10 可知, 该部分学生人数 = 1 + 5 + 2 + 4 + 10 = 22.

(d) 由表 14.11 可知, 两门课都不及格的学生人数为 3 + 3 + 6 + 5 = 17. 故至少一门及格的学生人数为 100 - 17 = 83, 由此可知百分比为: 83/100 = 83%.

表 14.9 有时称为二元频数表, 或二元频数分布. 表中的每一方格称为一个单元, 相应于一对分组区间. 单元中的数值称为单元频数. 例如, (a) 中数 4 即为对应于数学分为 70~79, 物理分为 80~89 的分组区间上的单元频数.

表 14.9 最后一行和最后一列的值称为边际和, 或边际频数. 它们分别对应于数学成绩和物理成绩的各自的频数分布.

表 14.9

		数学成绩						
		40~49	50~59	60~69	70~79	80~89	90~99	和
物 理 成 绩	90~99				2	4	4	10
	80~89			1	4	6	5	16
	70~79			5	10	8	1	24
	60~69	1	4	9	5	2		21
	50~59	3	6	6	2			17
	40~49	3	5	4				12
	和	7	15	25	23	20	10	100

表 14.10

		数学成绩	
		60~69	70~79
物理成绩	90~99		2
	80~89	1	4
	70~79	5	10

表 14.11

		数学成绩	
		40~49	50~59
物理成绩	50~59	3	6
	40~49	3	5

14.18 对二元频数表中的分组数据(表 14.9),如何修正习题 14.15 中的公式.

解 对分组数据,我们可以把变量 X 与 Y 的不同值作为分类数值,而 f_X 与 f_Y 为相应的分类频数,记在二元频数表的最后一行和最后一列.若设 f 为相应分类数值对 (X, Y) 的不同单元的单元频数,则我们可以把习题 14.15 中的公式变为

$$r = \frac{N \sum fXY - (\sum f_X X)(\sum f_Y Y)}{\sqrt{[N \sum f_X X^2 - (\sum f_X X)^2] \cdot [N \sum f_Y Y^2 - (\sum f_Y Y)^2]}} \quad (35)$$

若设 $X = A + c_X u_X$, $Y = B + c_Y u_Y$, 其中 c_X 和 c_Y 为分组区间的组距(假设为常数), A 和 B 为对应变量的任意分类成绩,公式(35)即变为本章的公式(21):

$$r = \frac{N \sum f u_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N \sum f_X u_X^2 - (\sum f_X u_X)^2] \cdot [N \sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}} \quad (21)$$

这是在前章节中为计算均值,标准差和高阶矩而采用的编码方法.

14.19 求习题 14.17 中数学和物理成绩的线性相关系数.

解 利用公式(21),相关的计算结果见表 14.12,称为**相关表**. $\sum f_X$, $\sum f_X u_X$, $\sum f_X u_X^2$, $\sum f_Y$, $\sum f_Y u_Y$, $\sum f_Y u_Y^2$ 通过编码方法计算.

表 14.12

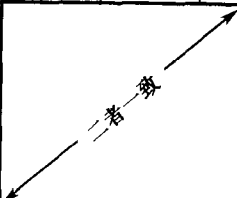
		数 学 成 绩										
		X	44.5	54.5	64.5	74.5	84.5					94.5
物理成绩	Y	u_X	-2	-1	0	1	2	3	f_Y	$f_Y u_Y$	$f_Y u_Y^2$	每行的角数和
	u_Y											
	94.5	2				2	4	4	10	20	40	44
	84.5	1			1	4	6	5	16	16	16	31
	74.5	0			5	10	8	1	24	0	0	0
	64.5	-1	1	4	9	5	2		21	-21	21	-3
	54.5	-2	3	6	6	2			17	-34	68	20
	44.5	-3	3	5	4				12	-36	108	33
f_X		7	15	25	23	20	10	$\sum f_X = \sum f_Y = N = 100$	$\sum f_Y u_Y = -55$	$\sum f_Y u_Y^2 = 253$	$\sum f_X u_X = 125$	
$f_X u_X$		-14	-15	0	23	40	30	$\sum f_X u_X = 64$				
$f_X u_X^2$		28	15	0	23	80	90	$\sum f_X u_X^2 = 236$				
每列的角数和		32	31	0	-1	24	39	$\sum f_X u_X u_Y = 125$				

表 14.12 中每一单元的右下角处的数表示乘积 fu_Xu_Y , 其中 f 是单元频数. 每一行的角边数之和列在同一行的最后一列; 而每一列的角边数之和列在同一列的最后一行. 最后一行与最后一列的和相等, 表示 $\sum fu_Xu_Y$ 的值.

从表 14.12, 有

$$r = \frac{N \sum fu_Xu_Y - \left(\sum f_Xu_X \right) \left(\sum f_Yu_Y \right)}{\sqrt{\left[N \sum f_Xu_X^2 - \left(\sum f_Xu_X \right)^2 \right] \cdot \left[N \sum f_Yu_Y^2 - \left(\sum f_Yu_Y \right)^2 \right]}}$$

$$= \frac{100 \times 125 - 64 \times (-55)}{\sqrt{(100 \times 236 - 64^2)(100 \times 253 - (-55)^2)}} = \frac{16020}{\sqrt{19504 \times 22275}} = 0.7686$$

14.20 用表 14.12 来计算 (a) s_X , (b) s_Y , (c) s_{XY} , 由此验证公式 $r = s_{XY} / (s_X \cdot s_Y)$.

解

$$(a) s_X = c_X \sqrt{\frac{\sum f_Xu_X^2}{N} - \left(\frac{\sum f_Xu_X}{N} \right)^2} = 10 \sqrt{\frac{236}{100} - \left(\frac{64}{100} \right)^2} = 13.966$$

$$(b) s_Y = c_Y \sqrt{\frac{\sum f_Yu_Y^2}{N} - \left(\frac{\sum f_Yu_Y}{N} \right)^2} = 10 \sqrt{\frac{253}{100} - \left(\frac{-55}{100} \right)^2} = 14.925$$

$$(c) s_{XY} = c_X c_Y \left[\frac{\sum fu_Xu_Y}{N} - \left(\frac{\sum f_Xu_X}{N} \right) \left(\frac{\sum f_Yu_Y}{N} \right) \right]$$

$$= 10 \times 10 \times \left[\frac{125}{100} - \left(\frac{64}{100} \right) \left(\frac{-55}{100} \right) \right] = 160.20$$

因此数学与物理成绩的标准差分别为 14.0 和 14.9, 而它们的协方差为 160.2. 因此相关系数为

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{160.2}{13.966 \times 14.925} = 0.7686$$

与习题 14.19 的结论一致.

回归直线和相关系数

14.21 证明 Y 关于 X 的回归直线方程与 X 关于 Y 的回归直线方程分别为:

$$(a) Y - \bar{Y} = \frac{rs_Y}{s_X} (X - \bar{X}), (b) X - \bar{X} = \frac{rs_X}{s_Y} (Y - \bar{Y}).$$

证明 (a) 从习题 13.15(a) 可知, Y 关于 X 的回归直线方程为

$$y = \frac{\sum xy}{\sum x^2} \cdot x \quad \text{或} \quad Y - \bar{Y} = \frac{\sum xy}{\sum x^2} (X - \bar{X})$$

又由习题 14.10 结论, 有

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

因此

$$\frac{\sum xy}{\sum x^2} = \frac{r \sqrt{(\sum x^2)(\sum y^2)}}{\sum x^2} = \frac{r \sqrt{\sum y^2}}{\sqrt{\sum x^2}} = \frac{rs_Y}{s_X}$$

即得所需结论.

(b) 交换(a)中 X 和 Y 的位置即得所需结论.

14.22 若 Y 关于 X 的回归直线和 X 关于 Y 的回归直线分别由 $Y = a_0 + a_1 X$, $X = b_0 + b_1 Y$ 给出. 证明 $a_1 b_1 = r^2$.

证明 由习题 14.21 的结论可知

$$a_1 = \frac{rs_Y}{s_X}, \quad b_1 = \frac{rs_X}{s_Y}$$

因此

$$a_1 b_1 = \frac{r s_Y}{s_X} \cdot \frac{r s_X}{s_Y} = r^2$$

这个结论可以作为线性相关系数定义的出发点.

14.23 用习题 14.22 的结论求习题 14.1 中数据的线性相关系数.

解 由习题 14.1 可知: $a_1 = 484/1016 = 0.476$, $b_1 = 484/467 = 1.036$, 因此 $r^2 = a_1 b_1 = (484/1016) \times (484/467)$, 得 $r = 0.7027$. 这和习题 14.9, 14.14 和 14.16 中得到的结论一致.

14.24 对习题 14.19 中的数据, 写出 (a) Y 关于 X 的回归直线方程, (b) X 关于 Y 的回归直线方程.

解 由习题 14.19 的相关表(表 14.12)得

$$\bar{X} = A + c_X \frac{\sum f_X u_X}{N} = 64.5 + \frac{10 \times 64}{100} = 70.9$$

$$\bar{Y} = B + c_Y \frac{\sum f_Y u_Y}{N} = 74.5 + \frac{10 \times (-55)}{100} = 69.0$$

由习题 14.20 的结论, 有 $s_X = 13.966$, $s_Y = 14.925$ 和 $r = 0.7686$. 现用习题 14.21(a) 与 (b) 的结论得到相应的回归方程.

$$(a) Y - \bar{Y} = \frac{r s_Y}{s_X} (X - \bar{X}),$$

$$Y - 69.0 = \frac{0.7686 \times 14.925}{13.966} (X - 70.9) = 0.821 (X - 70.9)$$

$$(b) X - \bar{X} = \frac{r s_X}{s_Y} (Y - \bar{Y}),$$

$$X - 70.9 = \frac{0.7686 \times 13.966}{14.925} (Y - 69.0) = 0.719 (Y - 69.0)$$

14.25 对习题 14.19 的数据, 用习题 14.20 的结论, 计算估计的标准误差 (a) $s_{Y \cdot X}$, (b) $s_{X \cdot Y}$.

解 (a) $s_{Y \cdot X} = s_Y \sqrt{1 - r^2} = 14.925 \times \sqrt{1 - 0.7686^2} = 9.548$

(b) $s_{X \cdot Y} = s_X \sqrt{1 - r^2} = 13.966 \times \sqrt{1 - 0.7686^2} = 8.934$

14.26 表 14.13 给出了美国在 1990 年到 1996 年消费者在食品和医疗保健方面的价格指数 (基于 1982~1984 年的价格, 均值为 100). 计算这两类指数的相关系数, 并给出这个系数的 Minitab 解.

表 14.13

年份	1990	1991	1992	1993	1994	1995	1996
食品	134.2	136.3	137.9	140.9	144.3	148.4	153.3
医疗保健	162.8	177.0	190.1	201.4	211.0	220.5	228.2

来源: 美国劳工统计局.

解 食品和医疗保健消费指数分别记为 X 和 Y , 一些相关计算见表 14.14.

表 14.14

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy	y^2
132.4	162.8	-9.53	-35.91	90.821	342.222	1289.53
136.3	177.0	-5.63	-21.71	31.697	122.227	471.32
137.9	190.1	-4.03	-8.61	16.241	34.698	74.13
140.9	201.4	-1.03	2.69	1.061	-2.771	7.24
144.3	211.0	2.37	12.29	5.617	29.127	151.04
148.4	220.5	6.47	21.79	41.861	140.981	474.80
153.3	228.2	11.37	29.49	129.277	335.301	869.66
$\bar{X} = 141.93$	$\bar{Y} = 198.71$			$\sum x^2 = 316.15$	$\sum xy = 1001.8$	$\sum y^2 = 3337.7$

然后利用积-矩公式可得

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{1001.8}{\sqrt{316.15 \times 3337.7}} = 0.975$$

把 X 值输入到 c1 列, Y 的值输入到 c2 列, 利用命令 correlation c1 c2 即可得到相关系数.

MTB>correlation c1 c2

Correlations (Pearson)

Correlation of X and Y = 0.975

非线性相关

14.27 对表 14.15 中数据拟合一条形如 $Y = a_0 + a_1X + a_2X^2$ 的最小二乘抛物线.

表 14.15

X	1.2	1.8	3.1	4.9	5.7	7.1	8.6	9.8
Y	4.5	5.9	7.0	7.8	7.2	6.8	4.5	2.7

解 第十三章中正规方程(23)为

$$\begin{aligned}\sum Y &= a_0N + a_1\sum X + a_2\sum X^2 \\ \sum XY &= a_0\sum X + a_1\sum X^2 + a_2\sum X^3 \\ \sum X^2Y &= a_0\sum X^2 + a_1\sum X^3 + a_2\sum X^4\end{aligned}\quad (36)$$

和的计算列在表 14.16 中, 因 $N=8$, (36)式即为

$$\begin{aligned}8a_0 + 42.2a_1 + 291.20a_2 &= 46.4 \\ 42.2a_0 + 291.20a_1 + 2275.35a_2 &= 230.42 \\ 291.20a_0 + 2275.35a_1 + 18971.92a_2 &= 1449.00\end{aligned}\quad (37)$$

解得 $a_0 = 2.588$, $a_1 = 2.065$, $a_2 = -0.2110$. 因此所求最小二乘回归抛物线方程为

$$Y = 2.588 + 2.065X - 0.2110X^2$$

表 14.16

X	Y	X^2	X^3	X^4	XY	X^2Y
1.2	4.5	1.44	1.73	2.08	5.40	6.48
1.8	5.9	3.24	5.83	10.49	10.62	19.12
3.1	7.0	9.61	29.79	92.35	21.70	67.27
4.9	7.8	24.01	117.65	576.48	38.22	187.28
5.7	7.2	32.49	185.19	1055.58	41.04	233.93
7.1	6.8	50.41	357.91	2541.16	48.28	342.79
8.6	4.5	73.96	636.06	5470.12	38.70	332.82
9.8	2.7	96.04	941.19	9223.66	26.46	259.31
$\sum X = 42.2$	$\sum Y = 46.4$	$\sum X^2 = 291.20$	$\sum X^3 = 2275.35$	$\sum X^4 = 18971.92$	$\sum XY = 230.42$	$\sum X^2Y = 1449.00$

14.28 利用习题 14.27 中的最小二乘抛物线来估计对应于给定 X 的 Y 值.

解 对 $X = 1.2$, $Y_{\text{est}} = 2.588 + 2.065 \times 1.2 - 0.2110 \times 1.2^2 = 4.762$. 其他值也可类似求出. 结果见表 14.17, 它也给出了 Y 的实际值.

表 14.17

Y_{est}	4.762	5.621	6.962	7.640	7.503	6.613	4.741	2.561
Y	4.5	5.9	7.0	7.8	7.2	6.8	4.5	2.7

- 14.29 (a) 求习题 14.27 中变量 X 与 Y 之间的线性相关系数.
 (b) 假设习题 14.27 中抛物线关系已知, 求变量间的非线性相关系数.
 (c) 解释(a), (b)所得相关系数间的差异.
 (d) 假定 X 与 Y 间存在抛物线关系. 求残差平方和与总变差的百分比.

解 (a) 用表 14.16 中已有的计算值及 $\sum Y^2 = 290.52$, 可得

$$r = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{[N \sum X^2 - (\sum X)^2] \cdot [N \sum Y^2 - (\sum Y)^2]}}$$

$$= \frac{8 \times 230.42 - 42.2 \times 46.4}{\sqrt{(8 \times 291.20 - 42.2^2)(8 \times 290.52 - 46.4^2)}} = -0.3743$$

(b) 由表 14.16 可知 $\bar{Y} = (\sum Y)/N = 46.4/8 = 5.80$; 因此总变差为 $\sum (Y - \bar{Y})^2 = 21.40$. 由表 14.17 知回归平方和为 $\sum (Y_{\text{est}} - \bar{Y})^2 = 21.02$, 因此

$$r^2 = \frac{\text{回归平方和}}{\text{总变差}} = \frac{21.02}{21.40} = 0.9822,$$

即

$$r = 0.9911 \quad \text{或} \quad r = 0.99$$

(c) 由(a)中给出的线性相关系数 -0.3743 , 仅能表明 X 与 Y 之间实际上没有线性关系. 然而由(b)中所得相关系数 0.99 可知, X 与 Y 之间存在非常好的非线性关系, 即抛物线关系.

(d) $\frac{\text{残差平方和}}{\text{总变差}} = 1 - r^2 = 1 - 0.9822 = 0.0178$

因此残差平方和占总变差的 1.78% , 这是由于某种随机波动或某种未考虑到的原因所引起的.

- 14.30 对习题 14.27 中的数据, 写出(a) s_Y , (b) $s_{Y \cdot X}$.

解 (a) 从习题 14.29(b)可知 $\sum (Y - \bar{Y})^2 = 21.40$, 故

$$s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}} = \sqrt{\frac{21.40}{8}} = 1.636 \quad \text{或} \quad 1.64$$

(b) 解法一 利用(a)的结果以及习题 14.29(b)的结论, 可得 Y 关于 X 的估计的标准误差

$$s_{Y \cdot X} = s_Y \sqrt{1 - r^2} = 1.636 \cdot \sqrt{1 - 0.9911^2} = 0.218 \quad \text{或} \quad 0.22$$

解法二 由习题 14.29, 有

$$s_{Y \cdot X} = \sqrt{\frac{\sum (Y - Y_{\text{est}})^2}{N}} = \sqrt{\frac{\text{残差平方和}}{N}}$$

$$= \sqrt{\frac{21.40 - 21.02}{8}} = 0.218 \quad \text{或} \quad 0.22$$

解法三 由习题 14.27 结论及 $\sum Y^2 = 290.52$ 有

$$s_{Y \cdot X} = \sqrt{\frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY - a_2 \sum X^2 Y}{N}} = 0.218 \quad \text{或} \quad 0.22$$

相关的抽样理论

- 14.31 容量为 18 的样本的相关系数为 0.32. 在显著性水平(a) 0.05, (b) 0.01 下, 是否可断定总体相关系数不为 0?

解 我们将在假设 $H_0: \rho = 0$ 和 $H_1: \rho > 0$ 之间进行选择.

$$t = \frac{r \sqrt{N-2}}{\sqrt{1-r^2}} = \frac{0.32 \cdot \sqrt{18-2}}{\sqrt{1-0.32^2}} = 1.35$$

(a) 在水平 0.05 下, 用 t 分布的单边检验, 若 $t > t_{0.95}(18-2) = 1.75$, 则拒绝 H_0 . 因为 $1.35 < 1.75$, 故在水平 0.05 下不能拒绝 H_0 .

(b) 既然在水平 0.05 下都不能拒绝 H_0 , 在水平 0.01 下就更不能拒绝 H_0 .

- 14.32 在显著性水平 0.05 下, 样本容量至少为多大, 才能认为相关系数 0.32 与 0 显著不同?

解 在水平 0.05 下用 t 分布的单边检验, 则 N 的最小值必须满足

$$\frac{0.32 \cdot \sqrt{N-2}}{\sqrt{1-0.32^2}} = t_{0.95}$$

自由度为 $N-2$. 当自由度无限大时, $t_{0.95} = 1.64$, 因此 $N = 25.6$.

对 $N = 26$: $\nu = 24$, $t_{0.95} = 1.71$, $t = 0.32 \sqrt{24} / \sqrt{1-0.32^2} = 1.65$

对 $N = 27$: $\nu = 25$, $t_{0.95} = 1.71$, $t = 0.32 \sqrt{25} / \sqrt{1-0.32^2} = 1.69$

对 $N = 28$: $\nu = 26$, $t_{0.95} = 1.71$, $t = 0.32 \sqrt{26} / \sqrt{1-0.32^2} = 1.72$

因此最小的样本容量应为 $N = 28$.

- 14.33 容量为 24 的样本的相关系数为 0.75. 在显著性水平 0.05 下, 是否能拒绝总体相关系数为 (a) $\rho = 0.60$, (b) $\rho = 0.50$ 的假设?

解 (a)

$$Z = 1.1513 \log \left(\frac{1+0.75}{1-0.75} \right) = 0.9730, \quad \mu_Z = 1.1513 \log \left(\frac{1+0.60}{1-0.60} \right) = 0.6932$$

且

$$\sigma_Z = \frac{1}{\sqrt{N-3}} = \frac{1}{\sqrt{21}} = 0.2182$$

因此

$$z = \frac{Z - \mu_Z}{\sigma_Z} = \frac{0.9730 - 0.6932}{0.2182} = 1.28$$

在显著性水平 0.05 下用正态分布的单边检验, 仅当 $z > 1.64$ 时拒绝假设. 因为 $z < 1.64$, 我们不能拒绝总体相关系数为 0.60 的假设.

(b) 若 $\rho = 0.50$, 则 $\mu_Z = 1.1513 \log 3 = 0.5493$ 以及 $z = (0.9730 - 0.5493) / 0.2182 = 1.94$. 因为 $z > 1.64$, 故在水平 0.05 下, 可以拒绝总体相关系数为 0.50 的假设.

- 14.34 21 名学生的数学与物理成绩间的相关系数为 0.80. 求该系数的 95% 的置信限.

解 因为 $r = 0.80$, $N = 21$, 故 μ_Z 的 95% 的置信限为

$$Z \pm 1.96 \sigma_Z = 1.1513 \log \left(\frac{1+r}{1-r} \right) \pm 1.96 \cdot \frac{1}{\sqrt{N-3}} = 1.0986 \pm 0.4620$$

因此 μ_Z 的 95% 的置信区间为 (0.5366, 1.5606). 若

$$\mu_Z = 1.1513 \log \left(\frac{1+\rho}{1-\rho} \right) = 0.5366, \text{ 则 } \rho = 0.4904$$

若

$$\mu_Z = 1.1513 \log \left(\frac{1+\rho}{1-\rho} \right) = 1.5606, \text{ 则 } \rho = 0.9155$$

故 ρ 的 95% 的置信限为 0.49 和 0.92.

- 14.35 从容量为 $N_1 = 28$ 和 $N_2 = 35$ 的两个样本中分别得出相应的相关系数 $r_1 = 0.50$, $r_2 = 0.30$. 在显著性水平 0.05 下判断两相关系数之间是否存在显著差异?

解

$$Z_1 = 1.1513 \log \frac{1+r_1}{1-r_1} = 0.5493, \quad Z_2 = 1.1513 \log \frac{1+r_2}{1-r_2} = 0.3095$$

且

$$\sigma_{Z_1-Z_2} = \sqrt{\frac{1}{N_1-3} + \frac{1}{N_2-3}} = 0.2669$$

我们将在假设 $H_0: \mu_{Z_1} = \mu_{Z_2}$ 和 $H_1: \mu_{Z_1} \neq \mu_{Z_2}$ 之间进行选择. 在假设 H_0 下

$$Z = \frac{Z_1 - Z_2 - (\mu_{Z_1} - \mu_{Z_2})}{\sigma_{Z_1-Z_2}} = \frac{0.5493 - 0.3095 - 0}{0.2669} = 0.8985$$

在水平 0.05 下用正态分布的双边检验, 若 $Z > 1.96$ 或 $Z < -1.96$, 则拒绝 H_0 . 故我们不能拒绝 H_0 , 即认为在水平 0.05 下, 两总体的相关系数之间没有显著差异.

回归的抽样理论

- 14.36** 在习题 14.1 中得知 Y 关于 X 的回归方程为 $Y = 35.82 + 0.476X$. 在显著性水平 0.05 下, 检验零假设: 总体回归方程的回归系数为 0.180, 其备择假设为: 回归系数大于 0.180. 用手算方法进行检验, 再用 Minitab 软件进行检验.

解

$$t = \frac{a_1 - A_1}{S_{Y \cdot X}/S_X} \sqrt{N-2} = \frac{0.476 - 0.180}{1.28/2.66} \sqrt{12-2} = 1.95$$

这是因为 $S_{Y \cdot X} = 1.28$ (见习题 14.5), $S_X = \sqrt{(\sum x^2)/N} = \sqrt{84.68/12} = 2.66$. 在显著性水平 0.05 下, 用 t 分布的单边检验, 若 $t > t_{0.95}(12-2) = 1.81$, 则拒绝回归系数为 0.180 的假设. 因此时 $t = 1.95 > 1.81$, 故拒绝零假设.

此题的 Minitab 输出如下:

MTB>Regress 'Y' 1 'X';

SUBC> Constant;

SUBC> Predict c7.

Regression Analysis

The regression equation is

$$Y = 35.8 + 0.476 X$$

Predictor	Coef	StDev	T	P
Constant	35.82	10.18	3.52	0.006
X	0.4764	0.1525	3.12	0.011

S = 1.404 R-Sq = 49.4% R-Sq(adj) = 44.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	19.214	19.214	9.75	0.011
Residual Error	10	19.703	1.970		
Total	11	38.917			

Predicted Values

Fit	StDev Fit	95.0% CI	95.0% PI
66.789	0.478	(65.724, 67.855)	(63.485, 70.094)
69.171	0.650	(67.723, 70.620)	(65.724, 72.618)

如下的部分给出假设检验所需的信息:

Predictor	Coef	StDev	T	P
Constant	35.82	10.18	3.52	0.006
X	0.4764	0.1525	3.12	0.011

检验统计量的值如下:

$$t = \frac{0.4764 - 0.180}{0.1525} = 1.94$$

输出的 t 值为 3.12, 用来检验回归系数为 0.180 的零假设. 为了检验回归系数的任何其他值, 需要进行上述类似的计算. 例如, 若对回归系数为 0.25 的假设进行检验, 则计算的 t 值等于

$$t = \frac{0.4764 - 0.25}{0.1525} = 1.48$$

回归系数为 0.25 的零假设将不会被拒绝.

- 14.37** 对习题 14.36 的回归系数求 95% 的置信限. 用手算方法进行计算, 再用 Minitab 软件进行检验.

解 置信区间可表示为

$$a_1 \pm \frac{t}{\sqrt{N-2}} \cdot \frac{S_{Y \cdot X}}{S_X}$$

因此 A_1 的 95% 置信限(通过设 $t = \pm t_{0.975}(12-2) = \pm 2.23$ 得到)为

$$a_1 \pm \frac{2.23}{\sqrt{12-2}} \cdot \frac{S_{Y \cdot X}}{S_X} = 0.476 \pm \frac{2.23}{\sqrt{10}} \cdot \frac{1.28}{2.66} = 0.476 \pm 0.340$$

即 A_1 的 95% 置信区间为(0.136, 0.816).

以下的 Minitab 输出从习题 14.36 得到, 它给出了计算 95% 置信区间所需的信息.

Predictor	Coef	StDev	T	P
Constant	35.82	10.18	3.52	0.006
X	0.4764	0.1525	3.12	0.011

有时称

$$\frac{1}{\sqrt{N-2}} \cdot \frac{S_{Y \cdot X}}{S_X}$$

为对应于回归系数的估计的标准误差. 输出中标准误差的值为 **0.1525**. 为了计算 95% 置信区间, 用 $t_{0.975}$ 乘以此标准误差, 然后再从 $a_1 = 0.476$ 中加上或减去此项, 从而得到 A_1 的置信区间如下

$$0.476 \pm 2.23 \times 0.1525 = 0.476 \pm 0.340$$

- 14.38** 在习题 14.1 中, 若父亲身高分别为(a) 65.0 英寸, (b) 70.0 英寸, 试写出儿子身高的 95% 置信限. 用手算方法进行计算, 再用 Minitab 软件进行计算.

解 因为 $t_{0.975}(12-2) = 2.23$, 则 Y_P 的 95% 置信限为

$$Y_0 \pm \frac{2.23}{\sqrt{N-2}} S_{Y \cdot X} \sqrt{N+1 + \frac{(X_0 - \bar{X})^2}{S_X^2}}$$

其中 $Y_0 = 35.82 + 0.476X_0$, $S_{Y \cdot X} = 1.28$, $S_X = 2.66$, $N = 12$.

(a) 若 $X_0 = 65.0$, 则 $Y_0 = 66.76$ 英寸. $(X_0 - \bar{X})^2 = (65.0 - 66.67)^2 = 2.78$. 因此 95% 置信限为

$$66.76 \pm \frac{2.23}{\sqrt{10}} \cdot 1.28 \cdot \sqrt{12+1 + \frac{2.78}{2.66^2}} = 66.76 \pm 3.30 \text{ 英寸}$$

即我们能以 95% 的概率保证儿子身高落在(63.46, 70.06)内.

(b) 若 $X_0 = 70.0$, 则 $Y_0 = 69.14$ 英寸. $(X_0 - \bar{X})^2 = (70.0 - 66.67)^2 = 11.09$. 因此 95% 置信限为 69.14 ± 3.45 英寸. 即我们能以 95% 的概率保证儿子身高落在(65.69, 72.59)内.

以下的 Minitab 输出从习题 14.36 得到, 它给出了计算 95% 置信区间所需的信息.

Predicted Values

Fit	StDev Fit	95.0% CI	95.0% PI
66.789	0.478	(65.724, 67.855)	(63.485, 70.094)
69.171	0.650	(67.723, 70.620)	(65.724, 72.618)

单个观测值的置信区间有时也认为是预测区间. 95% 的预测区间用黑体字表示. 不考虑舍入误差, 这些区间与上面计算结果一致.

- 14.39** 在习题 14.1 中, 若父亲身高分别为(a) 65.0 英寸, (b) 70.0 英寸, 试写出儿子平均身高的 95% 置信限. 用手算方法进行计算, 再用 Minitab 软件进行计算.

解 因为 $t_{0.975}(12-2) = 2.23$, 则 \bar{Y}_P 的 95% 置信限为

$$Y_0 \pm \frac{2.23}{\sqrt{10}} S_{Y \cdot X} \sqrt{1 + \frac{(X_0 - \bar{X})^2}{S_X^2}}$$

其中 $Y_0 = 35.82 + 0.476X_0$, $S_{Y \cdot X} = 1.28$, $S_X = 2.66$.

(a) 若 $X_0 = 65.0$, 则其置信限为 66.76 ± 1.07 或 65.7 和 67.8.

(b) 若 $X_0 = 70.0$, 则其置信限为 69.14 ± 1.45 或 67.7 和 70.6.

以下的 Minitab 输出从习题 14.36 得到, 它给出了 平均身高的 95% 置信区间, 用黑体字表示.

Predicted Values

Fit	StDev Fit	95.0% CI	95.0% PI
66.789	0.478	(65.724, 67.855)	(63.485, 70.094)
69.171	0.650	(67.723, 70.620)	(65.724, 72.618)

补充习题

线性回归和相关

14.40 表 14.18 给出了 10 个学生在生物学方面的两次测试的成绩(分别记为 X 和 Y).

- (a) 构造散点图;
- (b) 求 Y 关于 X 的最小二乘回归直线;
- (c) 求 X 关于 Y 的最小二乘回归直线;
- (d) 在(a)中散点图上画出(b)、(c)中的直线.

14.41 对表 14.18 中的数据, 求(a) $s_{Y \cdot X}$, (b) $s_{X \cdot Y}$.

表 14.18

第一次测验成绩(X)	6	5	8	8	7	6	10	4	9	7
第二次测验成绩(Y)	8	7	7	10	5	8	10	6	8	6

14.42 对习题 14.40 中的数据, 计算(a) Y 的总变差, (b) Y 的残差平方和, (c) Y 的回归平方和.

14.43 用习题 14.42 结论, 求习题 14.40 中两次成绩间的相关系数.

14.44 (a) 用积-矩公式求习题 14.40 中两次成绩间的相关系数. 并与习题 14.43 的结论比较.

(b) 直接从习题 14.40 的(b)、(c)的两回归直线的斜率求相关系数.

14.45 (a) 直接, (b) 用公式 $s_{XY} = r s_X s_Y$ 以及习题 14.43 或习题 14.44 的结论, 求习题 14.40 中数据的协方差.

14.46 表 14.19 给出了 12 名妇女的年龄 X 和血压 Y 的数据.

- (a) 求 X 与 Y 间的相关系数.
- (b) 求 Y 关于 X 的最小二乘回归方程.
- (c) 若一妇女年龄为 45 岁, 估计其血压.

表 14.19

年龄(X)	56	42	72	36	63	47	55	49	38	42	68	60
血压(Y)	147	125	160	118	149	128	150	145	115	140	152	155

14.47 对(a) 习题 13.32, (b) 习题 13.35 中数据分别求相关系数.

14.48 两变量 X 与 Y 的相关系数为 $r = 0.60$. 若 $s_X = 1.50$, $s_Y = 2.00$, $\bar{X} = 10$, $\bar{Y} = 20$, 求(a) Y 关于 X 的回归直线方程, (b) X 关于 Y 的回归直线方程.

14.49 对习题 14.48 中的数据, 求(a) $s_{Y \cdot X}$, (b) $s_{X \cdot Y}$.

14.50 若 $s_{Y \cdot X} = 3$, $s_Y = 5$, 求 r .

14.51 若 X 与 Y 的相关系数为 0.50, 通过回归方程求残差平方和占总变差的百分比.

14.52 (a) 证明 Y 关于 X 的回归直线方程可写成

$$Y - \bar{Y} = \frac{s_{XY}}{s_X^2}(X - \bar{X})$$

(b) 写出 X 关于 Y 的类似的回归直线方程.

14.53 (a) 求表 14.20 中给出的 X 与 Y 的相关系数

表 14.20

X	2	4	5	6	8	11
Y	18	12	10	8	7	5

(b) 对表 14.20 中的数据作变换: $X' = 2X + 6$, $Y' = 3X - 15$. 求 X' 与 Y' 间的相关系数, 并解释无论做

不做这样的变换, 所得结果始终与(a)中结果相同这一现象.

14.54 (a) 对习题 14.53 中(a)、(b)的数据, 分别求 Y 关于 X 的回归方程.

(b) 讨论回归方程间的关系.

14.55 (a) 证明 X 与 Y 的相关系数可写成

$$r = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sqrt{(X^2 - \bar{X}^2)(Y^2 - \bar{Y}^2)}}$$

(b) 用此方法, 求解习题 14.1.

14.56 证明相关系数与变量初始值及它们的单位的选择均无关. (提示: 假定 $X' = c_1X + A$, $Y' = c_2Y + B$, 其中 c_1, c_2, A 和 B 都是任意常数, 证明 X' 与 Y' 间的相关系数等于 X 与 Y 间的相关系数)

14.57 (a) 证明: 对线性回归来说

$$\frac{s_{Y.X}^2}{s_Y^2} = \frac{s_{X.Y}^2}{s_X^2}$$

始终成立.

(b), (a) 中结论对非线性回归也成立吗?

分组数据的相关系数

14.58 表 14.21 给出了 300 个美国成年男子身高与体重的频数表. 求身高与体重间的相关系数.

表 14.21

		身高 X (英寸)				
		59~62	63~66	67~70	71~74	75~78
体 重 Y (磅)	90~109	2	1			
	110~129	7	8	4	2	
	130~149	5	15	22	7	1
	150~169	2	12	63	19	5
	170~189		7	28	32	12
	190~209		2	10	20	7
	210~229			1	4	2

14.59 (a) 对习题 14.58 中的数据, 求 Y 关于 X 的最小二乘回归方程.

(b) 两男子身高分别为 64 和 72 英寸, 估计他们的体重.

14.60 对习题 14.58 中的数据, 求(a) $s_{Y.X}$, (b) $s_{X.Y}$.

14.61 对分组数据的相关系数, 建立本章的公式(21).

时间序列相关

14.62 表 14.22 给出了 1988 年到 1995 年每一消费者的年平均医疗保健支出和资产收入. 求二者间的相关系数.

表 14.22

年份	1988	1989	1990	1991	1992	1993	1994	1995
医疗保健支出	1298	1407	1480	1554	1634	1776	1755	1732
资产收入	17076	18194	19220	19715	20660	21288	22104	23233

来源: 劳工统计局和美国经济分析局

14.63 表 14.23 给出了 1989 年到 1998 年间每年 7 月份的平均气温和降雨量. 求二者的相关系数.

表 14.23

年份	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
温度(F)	78.1	71.8	75.6	72.7	75.3	73.6	75.1	75.3	73.8	70.4
降雨量(英寸)	6.23	3.64	3.42	2.84	1.83	2.82	4.04	2.56	1.18	4.19

相关的抽样理论

- 14.64** 容量为 27 的样本的相关系数为 0.40. 在显著性水平(a) 0.05, (b) 0.01 下判断, 相应总体的相关系数是否为 0?
- 14.65** 容量为 35 的样本的相关系数为 0.40. 在显著性水平 0.05 下能否拒绝假设: (a) 总体相关系数为 $\rho = 0.30$, (b) 总体相关系数为 $\rho = 0.70$?
- 14.66** 容量为 28 的样本的相关系数为 0.40. 求(a)总体相关系数的 95% 置信限, (b) 总体相关系数的 99% 置信限.
- 14.67** 若样本容量为 52, 求解习题 14.66.
- 14.68** 求(a) 习题 14.46, (b) 习题 14.58 中相关系数的 95% 的置信限.
- 14.69** 容量为 23 和 28 的样本的相关系数分别为 0.80 和 0.95. 在显著性水平(a) 0.05, (b) 0.01 下, 能否断定两系数间存在显著差异?

回归的抽样理论

- 14.70** 样本容量为 27 的样本中 Y 关于 X 的回归方程为 $Y = 25.0 + 2.00X$. 若 $s_{Y \cdot X} = 1.50$, $s_X = 3.00$, $\bar{X} = 7.50$, 求回归系数的(a) 95%, (b) 99% 的置信限.
- 14.71** 对习题 14.70, 在显著性水平 0.01 下, 对假设: 总体回归系数为(a) 1.70 和(b) 2.20 进行检验.
- 14.72** 对习题 14.70, 当 $X = 6.00$ 时, 求 Y 的 (a) 95%, (b) 99% 的置信限.
- 14.73** 对习题 14.70, 当 $X = 6.00$ 时, 求 Y 的平均值的 (a) 95%, (b) 99% 的置信限.
- 14.74** 对习题 14.46, 分别写出(a) Y 关于 X 的回归系数, (b) 45 岁妇女血压, (c) 45 岁妇女血压平均值的 95% 的置信限.

第十五章 多重相关与偏相关

多重相关

3 个或更多变量间存在的关系称为**多重相关**. 涉及多重相关问题的基本原理与在第十四章中所讨论过的简单相关的基本原理类似.

下标记号

为便于推广多变量问题, 有必要用到下标记号.

设 X_1, X_2, X_3, \dots 为考虑中的变量, 再令 $X_{11}, X_{12}, X_{13}, \dots$ 为相应变量 X_1 的值, $X_{21}, X_{22}, X_{23}, \dots$ 为相应变量 X_2 的值等等. 利用这些记号, 则和式 $X_{21} + X_{22} + \dots + X_{2N}$ 可写成 $\sum_{j=1}^N X_{2j}$, $\sum_j X_{2j}$ 或简记为 $\sum X_2$. 在不引起混淆的情况下, 我们用最后一个记号. 此时 X_2 的均值可写为: $\overline{X_2} = \sum X_2 / N$.

回归方程和回归平面

回归方程是为估计某因变量值的方程, 例如从自变量 X_2, X_3, \dots 估计 X_1 的值的方程称为 X_1 关于 X_2, X_3, \dots 的**回归方程**. 用函数记号, 有时可以写成 $X_1 = F(X_2, X_3, \dots)$ (读成 X_1 是 X_2, X_3 等的函数).

对三变量情形, X_1 关于 X_2, X_3 的最简单回归方程有以下形式

$$X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3 \quad (1)$$

其中 $b_{1.23}$, $b_{12.3}$ 和 $b_{13.2}$ 是常数. 如果我们在方程(1)中保持 X_3 为常数, 则 X_1 与 X_2 的关系曲线图是一斜率为 $b_{12.3}$ 的直线; 如果保持 X_2 为常数, 则 X_1 与 X_3 的关系曲线图是一斜率为 $b_{13.2}$ 的直线. 很清楚点后面的下标显示了每一情形中为常数的变量的下标.

因为 X_1 随着 X_2 或 X_3 的变化而有部分变化, 故我们称 $b_{12.3}$ 和 $b_{13.2}$ 分别为保持 X_3 为常数时 X_1 关于 X_2 的**偏相关系数**或保持 X_2 为常数时 X_1 关于 X_3 的**偏回归系数**.

方程(1)称为 X_1 关于 X_2, X_3 的**线性回归方程**. 在三维直角坐标系中, 它表示一平面, 称之为**回归平面**, 是第十三章中所讨论的两变量回归直线的推广.

最小二乘回归平面的正规方程

正如存在一拟合二维散点图中 N 个数据点集 (X, Y) 的最小二乘回归直线, 也存在一拟合三维散点图中 N 个数据点集 (X_1, X_2, X_3) 的最小二乘回归平面.

X_1 关于 X_2, X_3 的最小二乘回归平面具有方程(1)形式, 其中 $b_{1.23}$, $b_{12.3}$ 和 $b_{13.2}$ 通过同时解下列**正规方程**来确定:

$$\begin{aligned} \sum X_1 &= b_{1.23}N + b_{12.3} \sum X_2 + b_{13.2} \sum X_3 \\ \sum X_1 X_2 &= b_{1.23} \sum X_2 + b_{12.3} \sum X_2^2 + b_{13.2} \sum X_2 X_3 \\ \sum X_1 X_3 &= b_{1.23} \sum X_3 + b_{12.3} \sum X_2 X_3 + b_{13.2} \sum X_3^2 \end{aligned} \quad (2)$$

这些方程通过对方程(1)两边依次乘 $1, X_2, X_3$ 再求和可得.

除非特别说明, 以后任何时候所涉及回归方程都是指最小二乘回归方程.

若 $x_1 = X_1 - \overline{X_1}$, $x_2 = X_2 - \overline{X_2}$, $x_3 = X_3 - \overline{X_3}$, 则 X_1 关于 X_2 和 X_3 的回归方程可简化为

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \quad (3)$$

其中 $b_{12.3}$ 和 $b_{13.2}$ 通过同时解下列方程而得:

$$\begin{aligned}\sum x_1 x_2 &= b_{12.3} \sum x_2^2 + b_{13.2} \sum x_2 x_3 \\ \sum x_1 x_3 &= b_{12.3} \sum x_2 x_3 + b_{13.2} \sum x_3^2\end{aligned}\quad (4)$$

这些等价于正规方程(2)的方程可通过对方程(3)两边依次乘以 x_2, x_3 再求和而得.

回归平面和相关系数

若变量 X_1 与 X_2, X_1 与 X_3, X_2 与 X_3 之间的线性相关系数按第十四章中的方法计算, 分别记为 r_{12}, r_{13} 和 r_{23} (有时称为**零次相关系数**), 则最小二乘回归平面具有方程

$$\frac{x_1}{s_1} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \right) \frac{x_2}{s_2} + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \right) \frac{x_3}{s_3} \quad (5)$$

其中 $x_1 = X_1 - \bar{X}_1, x_2 = X_2 - \bar{X}_2, x_3 = X_3 - \bar{X}_3, s_1, s_2, s_3$ 分别为 X_1, X_2 和 X_3 的标准差(见习题 15.9).

注意, 若变量 X_3 不存在, 令 $X_1 = Y, X_2 = X$, 则方程(5)就转换为第十四章中的方程(25).

估计的标准误差

通过第十四章(8)式的推广, 我们可以定义 X_1 关于 X_2, X_3 的估计的标准误差为

$$s_{1.23} = \sqrt{\frac{\sum (X_1 - X_{1, \text{est}})^2}{N}} \quad (6)$$

其中 $X_{1, \text{est}}$ 是由回归方程(1)或(5)计算得到的 X_1 的估计值.

根据相关系数 r_{12}, r_{13} 和 r_{23} , 估计的标准误差也可由下式计算:

$$s_{1.23} = s_1 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \quad (7)$$

当样本容量 N 很大时, 第十四章中所讨论过的两变量的估计的标准误差的有关说明, 可通过用平行于回归平面的平面代替平行于回归直线的直线而推广到三维(即三变量)情形. 估计的标准误差的一个较佳形式可表为

$$\hat{s}_{1.23} = \sqrt{N/(N-3)} s_{1.23}$$

多重相关系数

通过推广第十四章中(12)或(14)式可定义多重相关系数. 例如, 对有两自变量的情形, 多重相关系数可由下式给出:

$$R_{1.23} = \sqrt{1 - \frac{s_{1.23}^2}{s_1^2}} \quad (8)$$

其中 s_1 是变量 X_1 的标准差, $s_{1.23}$ 由(6)或(7)式给出, 而数 $R_{1.23}^2$ 称为**多重判定系数**.

对于线性回归方程, 多重相关系数称为**线性多重相关系数**. 除非特别说明, 我们所说的多重相关都是指线性多重相关.

根据 r_{12}, r_{13} 和 r_{23} , (8)式可写成

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \quad (9)$$

多重相关系数, 如 $R_{1.23}$, 取 0 与 1 之间的值. 系数与 1 越接近, 则变量间的线性关系越好; 越接近 0, 线性关系则越差. 若多重相关系数等于 1, 则该相关称为**完全相关**. 尽管相关系数为 0 表明变量间没有线性关系, 但还可能存在某种**非线性关系**.

因变量的转换

当 X_1 作为因变量时, 上述结论都成立. 然而, 若我们把 X_3 而不是 X_1 作为因变量时, 我们只需将已得结论中的下标 1 换成 3, 3 换成 1 即可. 例如由方程(5), X_3 关于 X_1 与 X_2 的回归方程可写成

$$\frac{x_3}{s_3} = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2} \right) \frac{x_2}{s_2} + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2} \right) \frac{x_1}{s_1} \quad (10)$$

这里用到结论:

$$r_{32} = r_{23}, \quad r_{31} = r_{13}, \quad r_{21} = r_{12}$$

多于三个变量的推广

多于 3 个变量时也可得到类似上述的结论. 例如, X_1 关于 X_2, X_3, X_4 的线性回归方程可写成

$$X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \quad (11)$$

表示四维空间中一超平面. 通过对方程(11)两边依次乘以 1, X_2, X_3 和 X_4 , 再求和即可得确定常数 $b_{1.234}, b_{12.34}, b_{13.24}, b_{14.23}$ 的正规方程, 再把这些确定的常数代入(11)式就得到了 X_1 关于 X_2, X_3 和 X_4 的最小二乘回归方程, 这个最小二乘回归方程可写成类似于方程(5)的形式 (见习题 15.41).

偏相关

通常度量因变量与某特定自变量间的相关程度是很重要的, 此时, 其他变量都保持恒定, 以消除其他变量的影响. 这个相关程度可以通过类似于第十四章(12)式来定义, 称之为**偏相关系数**. 只是我们必须考虑在有与没有这个特定自变量的两种情况下所引出的回归平方和与残差平方和.

若我们用 $r_{12.3}$ 记为保持 X_3 恒定时, X_1 与 X_2 之间的偏相关系数, 则有

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \quad (12)$$

同样地, 若保持 X_3, X_4 恒定, X_1 与 X_2 之间的偏相关系数记为 $r_{12.34}$, 则有

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1 - r_{13.4}^2)(1 - r_{23.4}^2)}} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}} \quad (13)$$

这些结论是非常有用的, 因为通过它们, 所有偏相关系数都可最终只依赖于相关系数 r_{12}, r_{23} 等等.

在两变量 X 与 Y 情形中, 若两回归直线有方程: $Y = a_0 + a_1X$ 及 $X = b_0 + b_1Y$, 我们可知 $r^2 = a_1b_1$ (见习题 14.22). 这个结论能够被推广. 例如, 若

$$X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \quad (14)$$

$$X_4 = b_{4.123} + b_{41.23}X_1 + b_{42.13}X_2 + b_{43.12}X_3 \quad (15)$$

分别为 X_1 关于 X_2, X_3, X_4 及 X_4 关于 X_1, X_2, X_3 的线性回归方程. 则有

$$r_{14.23}^2 = b_{14.23}b_{41.23} \quad (16)$$

(见习题 15.18). 这可以作为线性偏相关系数定义的出发点.

多重相关系数与偏相关系数之间的关系

关于多重相关系数, 我们能找到一些有趣的结论. 例如, 我们有

$$1 - R_{1.23}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2) \quad (17)$$

$$1 - R_{1.234}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2) \quad (18)$$

这些结论的推广很容易得到.

非线性多重回归

关于线性多重回归的上述结论能推广到非线性多重回归情形, 此时多重相关系数与偏相关系数也可用类似上述方法来定义.

习题及解答

涉及三个变量的回归方程

15.1 用一适当的下标记号写出(a) X_2 关于 X_1 和 X_3 , (b) X_3 关于 X_1, X_2 和 X_4 , (c) X_5 关于 X_1, X_2, X_3 和 X_4 的回归方程.

解 (a) $X_2 = b_{2.13} + b_{21.3}X_1 + b_{23.1}X_3$

(b) $X_3 = b_{3.124} + b_{31.24}X_1 + b_{32.14}X_2 + b_{34.12}X_4$

(c) $X_5 = b_{5.1234} + b_{51.234}X_1 + b_{52.134}X_2 + b_{53.124}X_3 + b_{54.123}X_4$

15.2 写出对应回归方程(a) $X_3 = b_{3.12} + b_{31.2}X_1 + b_{32.1}X_2$, (b) $X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$ 的正规方程.

解 (a) 相继用 1, X_1 和 X_2 乘方程两边, 并求和即得正规方程

$$\sum X_3 = b_{3.12}N + b_{31.2} \sum X_1 + b_{32.1} \sum X_2$$

$$\sum X_1 X_3 = b_{3.12} \sum X_1 + b_{31.2} \sum X_1^2 + b_{32.1} \sum X_1 X_2$$

$$\sum X_2 X_3 = b_{3.12} \sum X_2 + b_{31.2} \sum X_1 X_2 + b_{32.1} \sum X_2^2$$

(b) 相继用 1, X_2, X_3 和 X_4 乘方程两边, 并求和即得正规方程

$$\sum X_1 = b_{1.234}N + b_{12.34} \sum X_2 + b_{13.24} \sum X_3 + b_{14.23} \sum X_4$$

$$\sum X_1 X_2 = b_{1.234} \sum X_2 + b_{12.34} \sum X_2^2 + b_{13.24} \sum X_2 X_3 + b_{14.23} \sum X_2 X_4$$

$$\sum X_1 X_3 = b_{1.234} \sum X_3 + b_{12.34} \sum X_2 X_3 + b_{13.24} \sum X_3^2 + b_{14.23} \sum X_3 X_4$$

$$\sum X_1 X_4 = b_{1.234} \sum X_4 + b_{12.34} \sum X_2 X_4 + b_{13.24} \sum X_3 X_4 + b_{14.23} \sum X_4^2$$

15.3 表 15.1 给出了 12 个男孩的体重 X_1 , 身高 X_2 , 年龄 X_3 .

(a) 求 X_1 关于 X_2 和 X_3 的最小二乘回归方程.

(b) 由给定的 X_2, X_3 的值确定 X_1 的值.

(c) 一男孩 9 岁, 身高 54 英寸, 估计其体重.

(d) 给出(a)的 Minitab 解.

表 15.1

体重(X_1)	64	71	53	67	55	58	77	57	56	51	76	68
身高(X_2)	57	59	49	62	51	50	55	48	52	42	61	57
年龄(X_3)	8	10	6	11	8	7	10	9	10	6	12	9

解 (a) X_1 关于 X_2 和 X_3 的线性回归方程可写成

$$X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

其正规方程为

$$\sum X_1 = b_{1.23}N + b_{12.3} \sum X_2 + b_{13.2} \sum X_3$$

$$\sum X_1 X_2 = b_{1.23} \sum X_2 + b_{12.3} \sum X_2^2 + b_{13.2} \sum X_2 X_3 \quad (19)$$

$$\sum X_1 X_3 = b_{1.23} \sum X_3 + b_{12.3} \sum X_2 X_3 + b_{13.2} \sum X_3^2$$

涉及和的一些计算由表 15.2 给出, 则上述正规方程为

表 15.2

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3
64	57	8	4096	3249	64	3648	512	456
71	59	10	5041	3481	100	4189	710	590
53	49	6	2809	2401	36	2597	318	294
67	62	11	4489	3844	121	4154	737	682
55	51	8	3025	2601	64	2805	440	408
58	50	7	3364	2500	49	2900	406	350
77	55	10	5929	3025	100	4235	770	550
57	48	9	3249	2304	81	2736	513	432
56	52	10	3136	2704	100	2912	560	520
51	42	6	2601	1764	36	2142	306	252
76	61	12	5776	3721	144	4636	912	732
68	57	9	4624	3249	81	3876	612	513
$\sum X_1$ = 753	$\sum X_2$ = 643	$\sum X_3$ = 106	$\sum X_1^2$ = 48139	$\sum X_2^2$ = 34843	$\sum X_3^2$ = 976	$\sum X_1X_2$ = 40830	$\sum X_1X_3$ = 6796	$\sum X_2X_3$ = 5779

$$\begin{aligned} 12b_{1.23} + 643b_{12.3} + 106b_{13.2} &= 753 \\ 643b_{1.23} + 34843b_{12.3} + 5779b_{13.2} &= 40830 \\ 106b_{1.23} + 5779b_{12.3} + 976b_{13.2} &= 6796 \end{aligned} \quad (20)$$

解得

$$b_{1.23} = 3.6512, b_{12.3} = 0.8546, b_{13.2} = 1.5063$$

故所求回归方程为

$$X_1 = 3.6512 + 0.8546X_2 + 1.5063X_3 \quad (21)$$

有避开解联立方程组的另一方法, 见习题 15.6.

(b) 利用回归方程(21), 通过代入相应的 X_2 和 X_3 的值, 求得 X_1 的估计值, 记为 $X_{1, \text{est}}$. 例如, 把 $X_2 = 57, X_3 = 8$ 代入(21)式, 即得 $X_{1, \text{est}} = 64.414$, 其他相应的 X_1 的估计值可以用相同方法得到, 见表 15.3.

表 15.3

$X_{1, \text{est}}$	64.414	69.136	54.564	73.206	59.286	56.925	65.717	58.229	63.153	48.582	73.857	65.920
X_1	64	71	53	67	55	58	77	57	56	51	76	68

(c) 把 $X_2 = 54, X_3 = 9$ 代入方程(21), 即得估计值 $X_{1, \text{est}} = 63.356$, 或大约为 63 磅.

(d) 体重记入第一列, 身高记入第二列, 年龄记入第三列, 命令 Regress 'Weight' on 2 predictors 'Height' and 'age'. 输出结果如下. 方程 $\text{Weight} = 3.7 + 0.855\text{Height} + 1.51\text{Age}$ 与上述得到的结果一致, 即 $X_1 = 3.65 + 0.855X_2 + 1.506X_3$.

MTB> Regress 'Weight' on 2 predictors 'Height' and 'Age'

The regression equation is

$$\text{Weight} = 3.7 + 0.855 \text{ Height} + 1.51 \text{ Age}$$

Predictor	Coef	StDev	T	P
Constant	3.65	16.17	0.23	0.826
Height	0.8546	0.4517	1.89	0.091
Age	1.506	1.414	1.07	0.315

$$S = 5.363 \quad R - \text{Sq} = 70.9\% \quad R - \text{Sq}(\text{adj}) = 64.4\%$$

15.4 对习题 15.3 中数据求标准差. (a) s_1 , (b) s_2 , (c) s_3 .

解 (a) s_1 是变量 X_1 的标准差, 则用表 15.2 中数据, 利用第四章方法, 得

$$s_1 = \sqrt{\frac{\sum X_1^2}{N} - \left(\frac{\sum X_1}{N}\right)^2} = \sqrt{\frac{48139}{12} - \left(\frac{753}{12}\right)^2} = 8.6035 \text{ 或 } 8.6 \text{ 磅}$$

(b)

$$s_2 = \sqrt{\frac{\sum X_2^2}{N} - \left(\frac{\sum X_2}{N}\right)^2} = \sqrt{\frac{34843}{12} - \left(\frac{643}{12}\right)^2} = 5.6930 \text{ 或 } 5.7 \text{ 英寸}$$

(c)

$$s_3 = \sqrt{\frac{\sum X_3^2}{N} - \left(\frac{\sum X_3}{N}\right)^2} = \sqrt{\frac{976}{12} - \left(\frac{106}{12}\right)^2} = 1.8181 \text{ 或 } 1.8 \text{ 年}$$

15.5 对习题 15.3 中数据, 求(a) r_{12} , (b) r_{13} , (c) r_{23} .

解 在(a)数 r_{12} 是变量 X_1 与 X_2 之间的线性相关系数, 则用第十四章中方法, 有

$$\begin{aligned} r_{12} &= \frac{N \sum X_1 X_2 - (\sum X_1)(\sum X_2)}{\sqrt{[N \sum X_1^2 - (\sum X_1)^2][N \sum X_2^2 - (\sum X_2)^2]}} \\ &= \frac{12 \times 40830 - 753 \times 643}{\sqrt{[12 \times 48139 - 753^2][12 \times 34843 - 643^2]}} = 0.8196 \text{ 或 } 0.82 \end{aligned}$$

(b)和(c) 用相应的公式, 可得 $r_{13} = 0.7698$ 或 0.77 , $r_{23} = 0.7984$ 或 0.80 .

15.6 用本章方程(5)及习题 15.4 和习题 15.5 的结论, 求习题 15.3(a).

解 在方程(5)两边乘以 s_1 , 则 X_1 关于 X_2, X_3 的回归方程变为

$$x_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \right) \left(\frac{s_1}{s_2} \right) x_2 + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \right) \left(\frac{s_1}{s_3} \right) x_3 \quad (22)$$

其中 $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$, $x_3 = X_3 - \bar{X}_3$. 用习题 15.4 及习题 15.5 的结论, 则方程(22)变为

$$x_1 = 0.8546x_2 + 1.5063x_3$$

因为 $\bar{X}_1 = \frac{\sum X_1}{N} = \frac{753}{12} = 62.750$, $\bar{X}_2 = \frac{\sum X_2}{N} = 53.583$, $\bar{X}_3 = 8.833$, 则所求方程可写为

$$X_1 - 62.750 = 0.8546(X_2 - 53.583) + 1.506(X_3 - 8.833)$$

与习题 15.3(a)中结论一致.

15.7 对习题 15.3 中数据求(a)相同年龄男孩, 再增加一英寸高度, 其相应的体重平均增量;
(b)相同身高男孩, 每年的体重平均增量.

解 由所得回归方程, 可解得(a)0.8546 或 0.9 磅, (b)1.5063 或 1.5 磅.

15.8 证明本章方程(3)和(4)由方程(1)和(2)所得.

解 对方程(2)中第一个式子两边同除以 N , 得

$$\bar{X}_1 = b_{1.23} + b_{12.3} \bar{X}_2 + b_{13.2} \bar{X}_3 \quad (23)$$

用方程(1)减去方程(23), 得

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

或

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \quad (24)$$

即为方程(3).

设 $X_1 = x_1 + \bar{X}_1$, $X_2 = x_2 + \bar{X}_2$, $X_3 = x_3 + \bar{X}_3$, 代入方程(2)的第二、三式, 通过代数式简化, 再由结论 $\sum x_1 = \sum x_2 = \sum x_3 = 0$, 则得

$$\begin{aligned} \sum x_1 x_2 &= b_{12.3} \sum x_2^2 + b_{13.2} \sum x_2 x_3 \\ &\quad + N \bar{X}_2 [b_{1.23} + b_{12.3} \bar{X}_2 + b_{13.2} \bar{X}_3 - \bar{X}_1] \end{aligned} \quad (25)$$

$$\begin{aligned} \sum x_1 x_3 &= b_{12.3} \sum x_2 x_3 + b_{13.2} \sum x_3^2 \\ &\quad + N \bar{X}_3 [b_{1.23} + b_{12.3} \bar{X}_2 + b_{13.2} \bar{X}_3 - \bar{X}_1] \end{aligned} \quad (26)$$

由方程(1)知(25)、(26)式等号右边方括号中的值为 0, 即得方程(4)式.

另一解法见习题 15.30.

15.9 建立方程(5), 即

$$\frac{x_1}{s_1} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \right) \left(\frac{x_2}{s_2} \right) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \right) \left(\frac{x_3}{s_3} \right) \quad (5)$$

解 由方程(25)、(26)得

$$\begin{aligned} b_{12.3} \sum x_2^2 + b_{13.2} \sum x_2 x_3 &= \sum x_1 x_2 \\ b_{12.3} \sum x_2 x_3 + b_{13.2} \sum x_3^2 &= \sum x_1 x_3 \end{aligned} \quad (27)$$

因为

$$s_2^2 = \frac{\sum x_2^2}{N}, \quad s_3^2 = \frac{\sum x_3^2}{N}$$

则

$$\sum x_2^2 = N s_2^2, \quad \sum x_3^2 = N s_3^2$$

又因为

$$r_{23} = \frac{\sum x_2 x_3}{\sqrt{(\sum x_2^2)(\sum x_3^2)}} = \frac{\sum x_2 x_3}{N s_2 s_3}$$

则

$$\sum x_2 x_3 = N s_2 s_3 r_{23}$$

同样有

$$\sum x_1 x_2 = N s_1 s_2 r_{12}, \quad \sum x_1 x_3 = N s_1 s_3 r_{13}$$

代入(27)式, 简化后得

$$\begin{aligned} b_{12.3} s_2 + b_{13.2} s_2 r_{23} &= s_1 r_{12} \\ b_{12.3} s_2 r_{23} + b_{13.2} s_3 &= s_1 r_{13} \end{aligned} \quad (28)$$

同时解方程(28)得

$$b_{12.3} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \right) \left(\frac{s_1}{s_2} \right), \quad b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \right) \left(\frac{s_1}{s_3} \right)$$

把这些代入方程 $x_1 = b_{12.3}x_2 + b_{13.2}x_3$, 再两边除以 s_1 即得结论.

估计的标准误差

15.10 对习题 15.3 中数据, 求 X_1 关于 X_2, X_3 的估计的标准误差.

解 由表 15.3 得

$$\begin{aligned} s_{1.23} &= \sqrt{\frac{\sum (X_1 - X_{1.est})^2}{N}} \\ &= \sqrt{\frac{(64 - 64.414)^2 + (71 - 69.136)^2 + \cdots + (68 - 65.920)^2}{12}} \\ &= 4.6447 \quad \text{或 } 4.6 \text{ 磅} \end{aligned}$$

修正的总体估计的标准误差为

$$\hat{s}_{1.23} = \sqrt{N(N-1)} s_{1.23} = 5.3 \text{ 磅}$$

15.11 用 $s_{1.23} = s_1 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$ 来求习题 15.10 的结论.

解 由习题 15.4(a)和 15.5 结论, 有

$$\begin{aligned} s_{1.23} &= 8.6035 \sqrt{\frac{1 - 0.8196^2 - 0.7698^2 - 0.7984^2 + 2 \times 0.8196 \times 0.7698 \times 0.7984}{1 - 0.7984^2}} \\ &= 4.6 \text{ 磅} \end{aligned}$$

注意, 用这种方法求估计的标准误差可以不用回归方程.

多重相关系数

15.12 对习题 15.3 中数据, 求 X_1 关于 X_2, X_3 的线性多重相关系数. 参照习题 15.3 中 Minitab 解来确定线性多重相关系数.

解 **解法一** 从习题 15.4(a) 和习题 15.10 的结论可得

$$R_{1.23} = \sqrt{1 - \frac{s_{1.23}^2}{s_1^2}} = \sqrt{1 - \frac{4.6447^2}{8.6035^2}} = 0.8418$$

解法二 由习题 15.5 的结论可得

$$\begin{aligned} R_{1.23} &= \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \\ &= \sqrt{\frac{0.8196^2 + 0.7698^2 - 2 \times 0.8196 \times 0.7698 \times 0.7984}{1 - 0.7984^2}} = 0.8468 \end{aligned}$$

注意, 多重相关系数 $R_{1.23}$ 比 r_{12} 或 r_{13} 都大 (见习题 15.5). 这始终是正确的, 也是可以料想的. 因为增加了有关自变量, 我们应该得到变量之间的较好的相关性.

在习题 15.3 中 Minitab 输出 $R_{1.23}^2 = 0.709$, 因此

$$R_{1.23} = \sqrt{0.709} = 0.802$$

15.13 对习题 15.3 中数据求 X_1 关于 X_2, X_3 的多重判定系数. 参照习题 15.3 的 Minitab 解来确定多重判定系数.

解 X_1 关于 X_2, X_3 的多重判定系数为

$$R_{1.23}^2 = (0.8418)^2 = 0.7086$$

由习题 15.12 中结论, 知 X_1 的变差约占总变差的 71%.

由习题 15.3 的 Minitab 输出可直接得到多重判定系数.

15.14 对习题 15.3 中数据, 计算 (a) $R_{2.13}$, (b) $R_{3.12}$, 并把它们与 $R_{1.23}$ 作比较.

解 (a)

$$\begin{aligned} R_{2.13} &= \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} \\ &= \sqrt{\frac{0.8196^2 + 0.7984^2 - 2 \times 0.8196 \times 0.7698 \times 0.7984}{1 - 0.7698^2}} = 0.8606 \end{aligned}$$

(b)

$$\begin{aligned} R_{3.12} &= \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}} \\ &= \sqrt{\frac{0.7698^2 + 0.7984^2 - 2 \times 0.8196 \times 0.7698 \times 0.7984}{1 - 0.8196^2}} = 0.8234 \end{aligned}$$

此例说明: 一般情况下, $R_{2.13}$, $R_{3.12}$ 和 $R_{1.23}$ 未必相等.

15.15 若 $R_{1.23} = 1$, 证明 (a) $R_{2.13} = 1$, (b) $R_{3.12} = 1$.

证明

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \quad (29)$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} \quad (30)$$

(a) 在 (29) 式中, 设 $R_{1.23} = 1$, 再对两边求平方, 得

$$r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23} = 1 - r_{23}^2$$

则有

$$r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23} = 1 - r_{13}^2$$

或

$$\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2} = 1$$

即 $R_{2,13}^2 = 1$ 或 $R_{2,13} = 1$, 因为多重相关系数非负.

(b) 同上, 可得 $R_{3,12} = 1$.

15.16 若 $R_{1,23} = 0$, 一定有 $R_{2,13} = 0$ 吗?

解 由(29)式可知 $R_{1,23} = 0$ 当且仅当

$$r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23} = 0 \text{ 或 } 2r_{12}r_{13}r_{23} = r_{12}^2 + r_{13}^2$$

则由(30)式有

$$R_{2,13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - (r_{12}^2 + r_{13}^2)}{1 - r_{13}^2}} = \sqrt{\frac{r_{23}^2 - r_{13}^2}{1 - r_{13}^2}}$$

即 $R_{2,13}$ 未必为 0.

偏相关

15.17 对习题 15.3 中数据, 求线性偏相关系数 (a) $r_{12,3}$, (b) $r_{13,2}$, (c) $r_{23,1}$.

解

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

$$r_{13,2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

$$r_{23,1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

利用习题 15.5 的结论可得 $r_{12,3} = 0.5334$, $r_{13,2} = 0.3346$, $r_{23,1} = 0.4580$. 这说明, 对相同年龄的男孩, 其体重与身高间的相关系数为 0.53; 对相同身高的男孩, 其年龄与体重间的相关系数为 0.33. 因为这些结论仅是由 12 个男孩的样本得出的, 当然没有大样本所得结果那么可靠.

15.18 若 $X_1 = b_{1,23} + b_{12,3}X_2 + b_{13,2}X_3$ 和 $X_3 = b_{3,12} + b_{32,1}X_2 + b_{31,2}X_1$ 分别为 X_1 关于 X_2, X_3 和 X_3 关于 X_2, X_1 的回归方程, 证明 $r_{13,2}^2 = b_{13,2}b_{31,2}$.

证明 X_1 关于 X_2, X_3 的回归方程可写成(见本章方程(5))

$$\begin{aligned} X_1 - \bar{X}_1 &= \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \right) \left(\frac{s_1}{s_2} \right) (X_2 - \bar{X}_2) \\ &\quad + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \right) \left(\frac{s_1}{s_3} \right) (X_3 - \bar{X}_3) \end{aligned} \quad (31)$$

而 X_3 关于 X_1, X_2 的回归方程可写成(见方程(10))

$$\begin{aligned} X_3 - \bar{X}_3 &= \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2} \right) \left(\frac{s_3}{s_2} \right) (X_2 - \bar{X}_2) \\ &\quad + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2} \right) \left(\frac{s_3}{s_1} \right) (X_1 - \bar{X}_1) \end{aligned} \quad (32)$$

由方程(31)和(32)可得 X_3 与 X_1 的系数分别为

$$b_{13,2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \right) \left(\frac{s_1}{s_3} \right), \quad b_{31,2} = \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2} \right) \left(\frac{s_3}{s_1} \right)$$

因此

$$b_{13,2}b_{31,2} = \frac{(r_{13} - r_{12}r_{23})^2}{(1 - r_{23}^2)(1 - r_{12}^2)} = r_{13,2}^2$$

15.19 若 $r_{12,3} = 0$, 证明 (a) $r_{13,2} = r_{13} \sqrt{\frac{1 - r_{23}^2}{1 - r_{12}^2}}$, (b) $r_{23,1} = r_{23} \sqrt{\frac{1 - r_{13}^2}{1 - r_{12}^2}}$.

证明 若 $r_{123} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} = 0$, 则有 $r_{12} = r_{13}r_{23}$.

$$\begin{aligned} (a) r_{13.2} &= \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} = \frac{r_{13} - (r_{13}r_{23})r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} \\ &= \frac{r_{13}(1-r_{23}^2)}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} = r_{13} \sqrt{\frac{1-r_{23}^2}{1-r_{12}^2}} \end{aligned}$$

(b) 交换(a)中 1 与 2 的位置即得(b)结论.

涉及四个或更多变量的多重或偏相关

15.20 一大学入学考试涉及三门课程: 数学、英语和常识. 为了通过考试效果来预测统计课程的成绩, 现对 200 名学生的数据进行收集和分析. 设 X_1 = 统计课程成绩, X_2 = 英语测试成绩, X_3 = 数学测试成绩, X_4 = 常识测试成绩. 由数据算得 $\bar{X}_1 = 75$, $s_1 = 10$, $\bar{X}_2 = 24$, $s_2 = 5$, $\bar{X}_3 = 15$, $s_3 = 3$, $\bar{X}_4 = 36$, $s_4 = 6$, $r_{12} = 0.90$, $r_{13} = 0.75$, $r_{14} = 0.80$, $r_{23} = 0.70$, $r_{24} = 0.70$, $r_{34} = 0.85$, 求 X_1 关于 X_2, X_3, X_4 的最小二乘回归方程.

解 推广习题 15.8 的结果, 则 X_1 关于 X_2, X_3, X_4 的最小二乘回归方程可写成

$$x_1 = b_{12.34}x_2 + b_{13.24}x_3 + b_{14.23}x_4 \quad (33)$$

其中 $b_{12.34}$, $b_{13.24}$ 和 $b_{14.23}$ 可由正规方程:

$$\begin{aligned} \sum x_1x_2 &= b_{12.34} \sum x_2^2 + b_{13.24} \sum x_2x_3 + b_{14.23} \sum x_2x_4 \\ \sum x_1x_3 &= b_{12.34} \sum x_2x_3 + b_{13.24} \sum x_3^2 + b_{14.23} \sum x_3x_4 \\ \sum x_1x_4 &= b_{12.34} \sum x_2x_4 + b_{13.24} \sum x_3x_4 + b_{14.23} \sum x_4^2 \end{aligned} \quad (34)$$

解得, 其中 $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$, $x_3 = X_3 - \bar{X}_3$, $x_4 = X_4 - \bar{X}_4$. 由数据算得

$$\begin{aligned} \sum x_2^2 &= Ns_2^2 = 5000 \\ \sum x_1x_2 &= Ns_1s_2r_{12} = 9000 \\ \sum x_2x_3 &= Ns_1s_3r_{23} = 2100 \\ \sum x_3^2 &= Ns_3^2 = 1800 \\ \sum x_1x_3 &= Ns_1s_3r_{13} = 4500 \\ \sum x_2x_4 &= Ns_2s_4r_{24} = 4200 \\ \sum x_4^2 &= Ns_4^2 = 7200 \\ \sum x_1x_4 &= Ns_1s_4r_{14} = 9600 \\ \sum x_3x_4 &= Ns_3s_4r_{34} = 3060 \end{aligned}$$

把以上结果代入方程(34)可得

$$b_{12.34} = 1.3333, \quad b_{13.24} = 0.0000, \quad b_{14.23} = 0.5556 \quad (35)$$

再代入方程(33), 则得所求回归方程

$$x_1 = 1.3333x_2 + 0.5556x_4$$

或

$$X_1 - 75 = 1.3333(X_2 - 24) + 0.5556(X_4 - 36) \quad (36)$$

即

$$X_1 = 22.9999 + 1.3333X_2 + 0.5556X_4$$

方程(34)的精确解为 $b_{12.34} = \frac{4}{3}$, $b_{13.24} = 0$, $b_{14.23} = \frac{5}{9}$, 故回归方程也可写成

$$X_1 = 23 + \frac{4}{3}X_2 + \frac{5}{9}X_4 \quad (37)$$

注意一个有趣现象: 回归方程中没有涉及 X_3 , 即英语成绩, 这并非说明统计成绩的预测不依赖于英语成绩. 相反, 这说明了统计成绩对英语分数的需要已充分体现到了其他分数当中.

- 15.21 在习题 15.20 中已知有两名学生的入学考试成绩分别为:
(a) 30, 18, 32; (b) 18, 20, 36. 问他们统计成绩预测为多少分?

解 (a) 把 $X_2 = 30, X_3 = 18, X_4 = 32$ 代入方程(37), 即得 $X_1 = 81$.
(b) 同理, 得 $X_1 = 67$.

- 15.22 对习题 15.20 中数据, 求偏相关系数 (a) $r_{12.34}$, (b) $r_{13.24}$, (c) $r_{14.23}$.

解 (a) 和 (b)

$$r_{12.34} = \frac{r_{12} - r_{14}r_{24}}{\sqrt{(1-r_{14}^2)(1-r_{24}^2)}}, \quad r_{13.24} = \frac{r_{13} - r_{14}r_{23}}{\sqrt{(1-r_{14}^2)(1-r_{23}^2)}}$$

$$r_{23.4} = \frac{r_{23} - r_{24}r_{34}}{\sqrt{(1-r_{24}^2)(1-r_{34}^2)}}$$

由习题 15.20 中数据, 可得 $r_{12.4} = 0.7935, r_{13.4} = 0.2215, r_{23.4} = 0.2791$. 因此

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1-r_{13.4}^2)(1-r_{23.4}^2)}} = 0.7814$$

$$r_{13.24} = \frac{r_{13.4} - r_{12.4}r_{23.4}}{\sqrt{(1-r_{12.4}^2)(1-r_{23.4}^2)}} = 0.0000$$

$$(c) r_{14.3} = \frac{r_{14} - r_{13}r_{24}}{\sqrt{(1-r_{13}^2)(1-r_{24}^2)}}, r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}, r_{24.3} = \frac{r_{24} - r_{23}r_{34}}{\sqrt{(1-r_{23}^2)(1-r_{34}^2)}}$$

由习题 15.20 数据, 得 $r_{14.3} = 0.4664, r_{12.3} = 0.7939, r_{24.3} = 0.2791$. 故得

$$r_{14.23} = \frac{r_{14.3} - r_{12.3}r_{24.3}}{\sqrt{(1-r_{12.3}^2)(1-r_{24.3}^2)}} = 0.4193$$

- 15.23 解释习题 15.22 中所得 (a) $r_{12.4}$, (b) $r_{13.4}$, (c) $r_{12.34}$, (d) $r_{14.3}$, (e) $r_{14.23}$.

解 (a) $r_{12.4} = 0.7935$ 代表有相同常识分数的学生的统计成绩与数学分数之间的线性相关系数, 其中不考虑学生的英语分数.

(b) $r_{13.4} = 0.2215$ 代表有相同常识分数的学生的统计成绩与英语分数之间的线性相关系数, 其中不考虑学生的数学分数.

(c) $r_{12.34} = 0.7814$, 表示有相同英语分数和相同常识分数的学生的统计成绩与数学分数之间的线性相关系数.

(d) $r_{14.3} = 0.4664$, 表示有相同英语分数的学生的统计成绩与常识分数间的线性相关系数, 其中不考虑学生的数学分数.

(e) $r_{14.23} = 0.4193$, 表示有相同数学分数和相同英语分数的学生的统计成绩与常识分数间的线性相关系数.

- 15.24 (a) 对习题 15.20 中数据验证

$$\frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1-r_{13.4}^2)(1-r_{23.4}^2)}} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1-r_{14.3}^2)(1-r_{24.3}^2)}} \quad (38)$$

(b) 解释 (a) 中等号明显成立.

解 (a) (38) 式左边的值等于 0.7814 (由习题 15.22(a) 可得), 再由 15.22 中结论可知 (38) 式右边的值也等于 0.7814, 故 (38) 式等号成立. 直接用代数运算也可证明一般情况下该等号成立.

(b) (38) 式左边为 $r_{12.34}$, 右边为 $r_{12.43}$, 而 $r_{12.34}$ 是保持 X_3, X_4 为常数, 变量 X_1 与 X_2 间的相关系数, $r_{12.43}$ 则是保持 X_4, X_3 为常数, 变量 X_1 与 X_2 间的相关系数, 由此明显得出等号成立.

- 15.25 对习题 15.20 中数据, 求 (a) 偏相关系数 $R_{1.234}$, (b) 估计 $s_{1.234}$ 的标准误差.

解 解法一 (a) $1 - R_{1.234}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2)$, 由习题 15.20 知 $r_{12} = 0.90$, 而由习题 15.22(c) 知 $r_{14.23} = 0.4193$,

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} = 0.3855$$

故得 $R_{1.234} = 0.9310$.

解法二 交换开头方程中下标 2 与 4 的位置, 则得

$$1 - R_{1.234}^2 = (1 - r_{14}^2)(1 - r_{13.4}^2)(1 - r_{12.34}^2) \text{ 或 } R_{1.234} = 0.9310,$$

其中直接利用习题 15.22(a) 中结论.

$$(b) R_{1.234} = \sqrt{1 - \frac{s_{1.234}^2}{s_1^2}} \text{ 或 } s_{1.234} = s_1 \sqrt{1 - R_{1.234}^2} = 10 \sqrt{1 - (0.9310)^2} = 3.650$$

补充习题

涉及三个变量的回归方程

15.26 用恰当的下标记号, 写出 (a) X_3 关于 X_1, X_2 , (b) X_4 关于 X_1, X_2, X_3 和 X_5 的回归方程.

15.27 写出 (a) X_2 关于 X_1, X_3 , (b) X_5 关于 X_1, X_2, X_3 和 X_4 的回归方程的正规方程.

15.28 表 15.4 给出了变量 X_1, X_2 和 X_3 的对应值.

(a) 求 X_3 关于 X_1, X_2 的最小二乘回归方程;

(b) 当 $X_1 = 10, X_2 = 6$ 时估计 X_3 的值.

表 15.4

X_1	3	5	6	8	12	14
X_2	16	10	7	4	3	2
X_3	90	72	54	42	30	12

15.29 一数学教师想测定期末考试成绩与该学期中两次小测验成绩的关系, 假定 X_1, X_2 和 X_3 分别表示学生第一次测验, 第二次测验和期末考试成绩. 他对 120 名学生的成绩统计分析得 $\bar{X}_1 = 6.8, \bar{X}_2 = 7.0, \bar{X}_3 = 74, s_1 = 1.0, s_2 = 0.80, s_3 = 9.0, r_{12} = 0.60, r_{13} = 0.70, r_{23} = 0.65$,

(a) 求 X_3 关于 X_1, X_2 的最小二乘回归方程.

(b) 已知两名学生的两次小测验成绩分别为 (1) 9, 7; (2) 4, 8, 估计这两名学生的期末成绩.

15.30 通过选择变量 X_2, X_3 使得 $\sum X_2 = \sum X_3 = 0$ 来解习题 15.8.

估计的标准误差

15.31 对习题 15.28 中数据, 求 X_3 关于 X_1, X_2 的估计的标准误差.

15.32 对习题 15.29 中数据, 求 (a) X_3 关于 X_1, X_2 , (b) X_1 关于 X_2, X_3 的估计的标准误差.

多重相关系数

15.33 对习题 15.28 中数据, 求 X_3 关于 X_1, X_2 的线性多重相关系数.

15.34 对习题 15.29 中数据, 计算 (a) $R_{3.12}$, (b) $R_{1.23}$, (c) $R_{2.13}$.

15.35 (a) 若 $r_{12} = r_{13} = r_{23} = r \neq 1$, 证明

$$R_{1.23} = R_{2.31} = R_{3.12} = \frac{r\sqrt{2}}{\sqrt{1+r}}$$

(b) $r = 1$ 时对 (a) 进行讨论.

15.36 若 $R_{1.23} = 0$, 证明 $|r_{23}| \geq |r_{12}|$ 且 $|r_{23}| \geq |r_{13}|$, 并对结论作出解释.

偏相关

15.37 对于习题 15.28 中的数据, 计算线性偏相关系数 (a) $r_{12.3}$, (b) $r_{13.2}$, (c) $r_{23.1}$, 并对结论作出解释.

15.38 对于习题 15.29 中的数据, 求解习题 15.37.

15.39 若 $r_{12} = r_{13} = r_{23} = r \neq 1$, 证明 $r_{12.3} = r_{13.2} = r_{23.1} = \frac{r}{1+r}$. 对 $r = 1$ 时进行讨论.

15.40 若 $r_{12.3} = 1$, 证明 (a) $|r_{13.2}| = 1$, (b) $|r_{23.1}| = 1$, (c) $R_{1.23} = 1$ 和 (d) $s_{1.23} = 0$.

涉及四个以上变量的多重相关和偏相关

15.41 证明 X_4 关于 X_1, X_2 和 X_3 的回归方程可写成

$$\frac{x_4}{s_4} = a_1 \left(\frac{x_1}{s_1} \right) + a_2 \left(\frac{x_2}{s_2} \right) + a_3 \left(\frac{x_3}{s_3} \right)$$

其中 a_1, a_2 和 a_3 由如下方程组给出:

$$a_1 r_{11} + a_2 r_{12} + a_3 r_{13} = r_{14}$$

$$a_1 r_{21} + a_2 r_{22} + a_3 r_{23} = r_{24}$$

$$a_1 r_{31} + a_2 r_{32} + a_3 r_{33} = r_{34}$$

且 $x_j = X_j - \bar{X}_j, j = 1, 2, 3, 4$. 推广到变量多于 4 个的情形.

- 15.42** 若 $\bar{X}_1 = 20, \bar{X}_2 = 36, \bar{X}_3 = 12, \bar{X}_4 = 80, s_1 = 1.0, s_2 = 2.0, s_3 = 1.5, s_4 = 6.0, r_{12} = -0.20, r_{13} = 0.40, r_{23} = 0.50, r_{14} = 0.40, r_{24} = 0.30$ 以及 $r_{34} = -0.10$, (a) 写出 X_4 关于 X_1, X_2 和 X_3 的回归方程, (b) 当 $X_1 = 15, X_2 = 40, X_3 = 14$ 时估计 X_4 .

- 15.43** 对习题 15.42 中的数据, 计算 (a) $r_{41.23}$, (b) $r_{42.13}$ 和 (c) $r_{43.12}$, 并对结论作出解释.

- 15.44** 对习题 15.42 中的数据, 计算 (a) $R_{4.123}$ 和 (b) $s_{4.123}$.

- 15.45** 某科学家收集了 4 个变量 T, U, V, W 的数据, 并认为变量间存在关系 $W = aT^b U^c V^d$, 其中 a, b, c, d 是未知常数. 通过上述等式, 他可以由 T, U, V 的值来估计 W . 设计一个程序来确定 a, b, c, d 的值 (提示: 可对等式两边取对数).

第十六章 方差分析

方差分析的目的

第八章我们通过抽样理论,在总体方差相同的假定下,对两总体均值差异的显著性进行了检验.在许多情况下有必要对 3 个或更多样本均值差异的显著性进行检验,或者等同于去检验零假设:样本均值全相等.

例 1 设在农业试验中,对 4 块土壤进行了不同的化学处理,小麦的平均产量分别为:28, 22, 18, 24(蒲式尔/英亩).那么这些均值之间是否有显著性差异,或者仅因为随机误差而造成观测值不一致?

这样的问题能通过运用 Fisher 提出的一个重要方法:方差分析得到解决,它主要利用第十一章已经讨论的 F 分布.

单向分类或单因素试验

在单因素试验中,获得 a 组独立的样本观察值(或观测值),每组所包含的观测值数目为 b .我们就说有 a 个处理或水平,每个处理重复 b 次.在例 1 中, $a = 4$.

单因素试验的结果以 a 行 b 列的列表表示,见表 16.1.此处 X_{jk} 表示第 j 行第 k 列的观测值, $j = 1, 2, \dots, a, k = 1, 2, \dots, b$.例如 X_{35} 表示第三个处理第五次试验的观测值.

表 16.1

处理 1	$X_{11}, X_{12}, \dots, X_{1b}$	$\bar{X}_{1.}$
处理 2	$X_{21}, X_{22}, \dots, X_{2b}$	$\bar{X}_{2.}$
\vdots	\vdots	\vdots
处理 a	$X_{a1}, X_{a2}, \dots, X_{ab}$	$\bar{X}_{a.}$

用 $\bar{X}_{j.}$ 表示第 j 行观测值的均值,则

$$\bar{X}_{j.} = \frac{1}{b} \sum_{k=1}^b X_{jk}, \quad j = 1, 2, \dots, a \quad (1)$$

$\bar{X}_{j.}$ 中的点表示对下标 k 已被求和. $\bar{X}_{j.}$ 被称之为组均值,处理均值或行均值.总均值即 a 个组所有观测值的平均值,用 \bar{X} 来表示,即

$$\bar{X} = \frac{1}{ab} \sum_{j=1}^a \sum_{k=1}^b X_{jk} \quad (2)$$

总变差,组内变差和组间变差

我们定义总变差 V 为每一观测值与总均值 \bar{X} 的离差平方和:

$$\text{总变差} = V = \sum_{j,k} (X_{jk} - \bar{X})^2 \quad (3)$$

对 $X_{jk} - \bar{X}$ 进行拆分

$$X_{jk} - \bar{X} = (X_{jk} - \bar{X}_{j.}) + (\bar{X}_{j.} - \bar{X}) \quad (4)$$

然后平方,并对 j, k 求和得到(见习题 16.1)

$$\sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 + \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 \quad (5)$$

或

$$\sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk} - \bar{X}_{j\cdot})^2 + b \sum_j (\bar{X}_{j\cdot} - \bar{X})^2 \quad (6)$$

我们把等式(5), (6)右边的第一个和式称为**组内变差**, 用 V_W 来表示它, 即

$$V_W = \sum_{j,k} (X_{jk} - \bar{X}_{j\cdot})^2 \quad (7)$$

等式(5), (6)右边的第二项称为**组间变差**, 用 V_B 表示, 即

$$V_B = \sum_{j,k} (\bar{X}_{j\cdot} - \bar{X})^2 = b \sum_j (\bar{X}_{j\cdot} - \bar{X})^2 \quad (8)$$

因此等式(5), (6)可以写成

$$V = V_W + V_B \quad (9)$$

计算变差的快捷方法

为了减小计算上述变差的工作量, 可以采用下面的形式来表示

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} \quad (10)$$

$$V_B = \frac{1}{b} \sum_j T_{j\cdot}^2 - \frac{T^2}{ab} \quad (11)$$

$$V_W = V - V_B \quad (12)$$

其中

$$T = \sum_{j,k} X_{jk}, \quad T_{j\cdot} = \sum_k X_{jk} \quad (13)$$

实际上, 从表上所有数据中各减去某个固定的值, 计算起来更方便, 且对最终的结果没有影响.

方差分析的数学模型

我们把表 16.1 的每一行看作是取自某一处理(或水平)所对应总体的容量为 b 的随机样本, X_{jk} 与第 j 种处理所对应的总体均值 μ_j 之间有一**随机误差或机会误差**, 用 ϵ_{jk} 来表示它, 则

$$X_{jk} = \mu_j + \epsilon_{jk} \quad (14)$$

这些误差被认为服从均值为 0, 方差为 σ^2 的正态分布, 令 $\mu = \frac{1}{a} \sum_{j=1}^a \mu_j$, $\alpha_j = \mu_j - \mu$, 则 $\mu_j = \mu + \alpha_j$. 等式(14)变为

$$X_{jk} = \mu + \alpha_j + \epsilon_{jk} \quad (15)$$

其中 $\sum_j \alpha_j = 0$ (见习题 16.9). 从等式(15)和假设 $\epsilon_{jk} \sim N(0, \sigma^2)$, 我们能推断出 X_{jk} 是服从均值为 $\mu + \alpha_j$, 方差为 σ^2 的正态分布的随机变量.

零假设 $H_0: \mu_1 = \mu_2 = \cdots = \mu_a$ 就等价于 $H_0: \alpha_j = 0, j = 1, 2, \cdots, a$, 或 $H_0: \mu_j = \mu, j = 1, 2, \cdots, a$. 假如 H_0 成立, a 个水平所对应的总体都应服从同一正态分布(即有相同的均值与方差). 这种情况下, 仅有一个总体(即所有处理从统计学角度来看都应是同一的); 换句话说, 处理间没有显著性差异.

变差的数学期望

从习题 16.10 可知, V_W, V_B, V 的数学期望如下:

$$E(V_W) = a(b-1)\sigma^2 \quad (16)$$

$$E(V_B) = (a-1)\sigma^2 + b \sum_j \alpha_j^2 \quad (17)$$

$$E(V) = (ab-1)\sigma^2 + b \sum_j \alpha_j^2 \quad (18)$$

从式(16)可知

$$E\left(\frac{V_w}{a(b-1)}\right) = \sigma^2 \quad (19)$$

因此无论 H_0 成立与否

$$\hat{S}_w^2 = \frac{V_w}{a(b-1)} \quad (20)$$

都是参数 σ^2 的一个最优(无偏)估计. 另一方面, 从等式(16), (18)可推断出, 仅当 H_0 成立(即 $\alpha_j = 0$)时, 才有

$$E\left(\frac{V_B}{a-1}\right) = \sigma^2 \quad E\left(\frac{V}{ab-1}\right) = \sigma^2 \quad (21)$$

因此仅当 H_0 成立时

$$\hat{S}_B^2 = \frac{V_B}{a-1} \quad \hat{S}^2 = \frac{V}{ab-1} \quad (22)$$

才是 σ^2 的无偏估计. 若 H_0 不成立, 从等式(17)可知

$$E(\hat{S}_B^2) = \sigma^2 + \frac{b}{a-1} \sum_j \alpha_j^2 \quad (23)$$

此时 \hat{S}_B^2 不是 σ^2 的无偏估计.

变差的分布

利用 χ^2 分布的附加性质, 我们能证明下面有关 V_w, V_B, V 分布的重要定理.

定理 1 $\frac{V_w}{\sigma^2} \sim \chi^2(a(b-1))$

定理 2 若 H_0 成立, 则 $\frac{V_B}{\sigma^2} \sim \chi^2(a-1), \frac{V}{\sigma^2} \sim \chi^2(ab-1)$.

注 无论 H_0 成立与否定理 1 都成立, 而定理 2 只有当 H_0 成立时才有效.

等均值零假设的 F 检验

若零假设 H_0 不成立(即各个处理的均值不全等), 从等式(23)我们能看出, \hat{S}_B^2 总比 σ^2 要大些, 且均值间的差异越大, 效应就变得越明显; 另一方面, 从等式(19), (20)能看出, 无论均值相等与否, \hat{S}_w^2 都可看作与 σ^2 是等同的, 因此就得到检验 H_0 的一个很好的统计量 $\frac{\hat{S}_B^2}{\hat{S}_w^2}$. 若这个统计量明显偏大, 我们即可推断, 各个处理均值之间有明显差异, 因此拒绝 H_0 ; 否则接受 H_0 , 再作进一步分析.

为了应用统计量 $\frac{\hat{S}_B^2}{\hat{S}_w^2}$, 我们必须了解它的抽样分布.

定理 3 统计量 $F = \frac{\hat{S}_B^2}{\hat{S}_w^2} \sim F(a-1, a(b-1))$.

定理 3 使我们能在某给定显著性水平下, 运用 F 分布的单边检验来检验零假设 H_0 (见第十一章).

方差分析表

上述检验的计算见表 16.2, 这样的表称为**方差分析表**. 实际上, 我们可通过(3)或(8)的长方法计算 V 和 V_B , 也可用(10)或(11)的短方法来计算, 然后利用 $V_w = V - V_B$ 计算 V_w . 注意, 总变差的自由度 $(ab-1)$ 等于组内变差与组间变差的自由度之和.

表 16.2

变 差	自由度	均 方	F
组间变差 $V_B = b \sum_j (\bar{X}_{j\cdot} - \bar{X})^2$	$a - 1$	$\hat{S}_B^2 = \frac{V_B}{a - 1}$	$\frac{\hat{S}_B^2}{\hat{S}_W^2}$ 服从自由度 $a - 1, a(b - 1)$ 的 F 分布
组内变差 $V_W = V - V_B$	$a(b - 1)$	$\hat{S}_W^2 = \frac{V_W}{a(b - 1)}$	
总变差 $V = V_B + V_W = \sum_{j,k} (X_{jk} - \bar{X})^2$	$ab - 1$		

观测值数目不等时所做的修正

假如处理 1, 处理 2……, 处理 a 所对应的样本容量不同, 分别为 N_1, N_2, \dots, N_a , 上述结论应被修正:

$$V = \sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} X_{jk}^2 - \frac{T^2}{N} \quad (24)$$

$$V_B = \sum_{j,k} (\bar{X}_{j\cdot} - \bar{X})^2 = \sum_j N_j (\bar{X}_{j\cdot} - \bar{X})^2 = \sum_j \frac{T_{j\cdot}^2}{N_j} - \frac{T^2}{N} \quad (25)$$

$$V_W = V - V_B \quad (26)$$

其中 $\sum_{j,k}$ 表示 j 不变, 对 k 从 1 到 N_j 求和, 然后再对 j 从 1 到 a 求和. 表 16.3 就是这种情况下的方差分析表.

表 16.3

变 差	自由度	均 方	F
组间变差 $V_B = \sum_j N_j (\bar{X}_{j\cdot} - \bar{X})^2$	$a - 1$	$\hat{S}_B^2 = \frac{V_B}{a - 1}$	$\frac{\hat{S}_B^2}{\hat{S}_W^2}$ 服从自由度为 $a - 1, N - a$ 的 F 分布
组内变差 $V_W = V - V_B$	$N - a$	$\hat{S}_W^2 = \frac{V_W}{N - a}$	
总变差 $V = V_B + V_W = \sum_{j,k} (X_{jk} - \bar{X})^2$	$N - 1$		

双向分类或双因素试验

单因素方差分析的思想可被推广到一般情况. 例 2 就是对双因素方差分析思想的一个阐述.

例 2 假设一个农业试验的目的是为了考察四种不同品种小麦的亩产量, 将每一品种种植在 5 块不同的土壤中, 因此共需 $4 \times 5 = 20$ 块地. 这种情况下, 若把几个小块合并成一个区组, 显然有利于观察. 因此我们把种植不同品种小麦的 4 个小块合并成一个区组, 这样就需要 5 个区组.

亩产量可能由于下面的原因而形成差异: (1) 小麦品种不同或 (2) 区组的不同. 所以此时存在着两个因素.

由例 2 的实验知, 我们通常会在试验中遇到两个因素: **处理**和**区组**, 一般简称为因素 1 和因素 2.

双因素试验的记号表示

假设有 a 个处理, b 个区组, 可构造表 16.4, 假设其中的每个符号均对应于一个观测值.

对处理 j 和区组 k , 可用 X_{jk} 表示其值, $\bar{X}_{j\cdot}$ ($j=1, 2, \dots, a$) 表示第 j 行数据的均值, $\bar{X}_{\cdot k}$ ($k=1, 2, \dots, b$) 表示第 k 列数据的均值, \bar{X} 表示总均值, 即

$$\bar{X}_{j\cdot} = \frac{1}{b} \sum_{k=1}^b X_{jk}, \quad \bar{X}_{\cdot k} = \frac{1}{a} \sum_{j=1}^a X_{jk}, \quad \bar{X} = \frac{1}{ab} \sum_{j,k} X_{jk} \quad (27)$$

表 16.4

	区 组				
	1	2	...	b	
处理 1	X_{11}	X_{12}	...	X_{1b}	$\bar{X}_{1\cdot}$
处理 2	X_{21}	X_{22}	...	X_{2b}	$\bar{X}_{2\cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
处理 a	X_{a1}	X_{a2}	...	X_{ab}	$\bar{X}_{a\cdot}$
	$\bar{X}_{\cdot 1}$	$\bar{X}_{\cdot 2}$...	$\bar{X}_{\cdot b}$	

双因素试验的变差

如同单因素试验一样, 我们可定义双因素试验的变差. 总变差(如(3)式)

$$V = \sum_{j,k} (X_{jk} - \bar{X})^2 \quad (28)$$

对 $X_{jk} - \bar{X}$ 进行拆分

$$X_{jk} - \bar{X} = (X_{jk} - \bar{X}_{j\cdot} - \bar{X}_{\cdot k} + \bar{X}) + (\bar{X}_{j\cdot} - \bar{X}) + (\bar{X}_{\cdot k} - \bar{X}) \quad (29)$$

然后平方, 并对 j, k 求和, 则

$$V = V_E + V_R + V_C \quad (30)$$

其中

$$V_E = \text{随机变差} = \sum_{j,k} (X_{jk} - \bar{X}_{j\cdot} - \bar{X}_{\cdot k} + \bar{X})^2$$

$$V_R = \text{处理间变差} = b \sum_{j=1}^a (\bar{X}_{j\cdot} - \bar{X})^2$$

$$V_C = \text{区组间变差} = a \sum_{k=1}^b (\bar{X}_{\cdot k} - \bar{X})^2$$

随机变差也称为残差或随机变差.

对应于(10), (11), (12)式, 我们也可得到计算 V, V_R, V_C, V_E 的快捷方法:

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} \quad (31)$$

$$V_R = \frac{1}{b} \sum_{j=1}^a T_{j\cdot}^2 - \frac{T^2}{ab} \quad (32)$$

$$V_C = \frac{1}{a} \sum_{k=1}^b T_{\cdot k}^2 - \frac{T^2}{ab} \quad (33)$$

$$V_E = V - V_R - V_C \quad (34)$$

其中

$$T = \sum_{j,k} X_{jk}, \quad T_{j\cdot} = \sum_{k=1}^b X_{jk}, \quad T_{\cdot k} = \sum_{j=1}^a X_{jk}$$

双因素方差分析

将(15)式所给出的单因素方差分析的数学模型进行推广, 可得到双因素试验的数学模型:

$$X_{jk} = \mu + \alpha_j + \beta_k + \epsilon_{jk} \quad (35)$$

其中 $\sum \alpha_j = 0, \sum \beta_k = 0$. μ 表示总体总均值, α_j 是对应 X_{jk} 中处理差异的部分(有时也称为处

理效应), β_k 是对应 X_{jk} 中区组差异的部分 (有时也称为区组效应). ϵ_{jk} 是对应 X_{jk} 中随机误差的部分, 如前所述, 假定 ϵ_{jk} 服从均值为 0, 方差为 σ^2 的正态分布, 则 X_{jk} 服从均值为 $\mu + \alpha_j + \beta_k$, 方差为 σ^2 的正态分布.

对应于 (16), (17), (18) 式, 可证明上述变差的数学期望

$$E(V_E) = (a-1)(b-1)\sigma^2 \quad (36)$$

$$E(V_R) = (a-1)\sigma^2 + b \sum_j \alpha_j^2 \quad (37)$$

$$E(V_C) = (b-1)\sigma^2 + a \sum_k \beta_k^2 \quad (38)$$

$$E(V) = (ab-1)\sigma^2 + b \sum_j \alpha_j^2 + a \sum_k \beta_k^2 \quad (39)$$

需要检验如下两个零假设:

$H_0^{(1)}$: 所有处理均值相等, 即 $\alpha_j = 0, j = 1, 2, \dots, a$.

$H_0^{(2)}$: 所有区组均值相等, 即 $\beta_k = 0, k = 1, 2, \dots, b$.

从 (36) 式可看出, $E(\hat{S}_E^2) = \sigma^2$, 因此不管假设 $H_0^{(1)}, H_0^{(2)}$ 成立与否, σ^2 都有一个无偏估计

$$\hat{S}_E^2 = \frac{V_E}{(a-1)(b-1)} \quad (40)$$

若假设 $H_0^{(1)}, H_0^{(2)}$ 成立, 则

$$\hat{S}_R^2 = \frac{V_R}{a-1}, \quad \hat{S}_C^2 = \frac{V_C}{b-1}, \quad \hat{S}^2 = \frac{V}{ab-1} \quad (41)$$

也均为 σ^2 的无偏估计. 若假设 $H_0^{(1)}, H_0^{(2)}$ 不成立, 从 (37), (38) 式可看出,

$$E(\hat{S}_R^2) = \sigma^2 + \frac{b}{a-1} \sum_j \alpha_j^2 \quad (42)$$

$$E(\hat{S}_C^2) = \sigma^2 + \frac{a}{b-1} \sum_k \beta_k^2 \quad (43)$$

下面定理与定理 1 和定理 2 类似:

定理 4 $\frac{V_E}{\sigma^2} \sim \chi^2((a-1)(b-1))$, $H_0^{(1)}, H_0^{(2)}$ 成立与否都适用.

定理 5 若 $H_0^{(1)}$ 成立, 则 $\frac{V_R}{\sigma^2} \sim \chi^2(a-1)$. 若 $H_0^{(2)}$ 成立, 则 $\frac{V_C}{\sigma^2} \sim \chi^2(b-1)$. 若 $H_0^{(1)}, H_0^{(2)}$ 均成立, 则 $\frac{V}{\sigma^2} \sim \chi^2(ab-1)$.

要检验 $H_0^{(1)}$, 自然联想到统计量 $\frac{\hat{S}_R^2}{\hat{S}_E^2}$, 因为从 (42) 式可看出, 如果处理均值差异显著, 则

\hat{S}_R^2 与 σ^2 间的差异也应该显著; 同理, 为了要检验 $H_0^{(2)}$, 就要考虑统计量 $\frac{\hat{S}_C^2}{\hat{S}_E^2}$ 和 $\frac{\hat{S}_R^2}{\hat{S}_E^2}$ 的分布见定理 6 (如同定理 3).

定理 6 若 $H_0^{(1)}$ 成立, 则统计量 $\frac{\hat{S}_R^2}{\hat{S}_E^2} \sim F(a-1, (a-1)(b-1))$. 若 $H_0^{(2)}$ 成立, 则统计量

$$\frac{\hat{S}_C^2}{\hat{S}_E^2} \sim F(b-1, (a-1)(b-1)).$$

由定理 6, 我们可在给定显著性水平下, 接受或拒绝 $H_0^{(1)}$ 或 $H_0^{(2)}$. 为了计算方便, 可仿照单因素方差分析, 得到双因素方差分析的方差分析表 (见表 16.5).

表 16.5

变 差	自由度	均 方	F
处理间变差 $V_R = b \sum_j (\bar{X}_{j.} - \bar{X})^2$	$a - 1$	$\hat{S}_R^2 = \frac{V_R}{a - 1}$	$\frac{\hat{S}_R^2}{\hat{S}_E^2} \sim F(a - 1, (a - 1)(b - 1))$
区组间变差 $V_C = a \sum_k (\bar{X}_{.k} - \bar{X})^2$	$b - 1$	$\hat{S}_C^2 = \frac{V_C}{b - 1}$	$\frac{\hat{S}_C^2}{\hat{S}_E^2} \sim F(b - 1, (a - 1)(b - 1))$
残差 $V_E = V - V_R - V_C$	$(a - 1)(b - 1)$	$\hat{S}_E^2 = \frac{V_E}{(a - 1)(b - 1)}$	
总变差 $V = V_R + V_C + V_E = \sum_{j,k} (X_{jk} - \bar{X})^2$	$ab - 1$		

有重复的双因素试验

在表 16.4 中, 相对于给定的处理与区组, 仅收集了一个观测值. 因素的更多信息经常通过重复试验得到, 即称为有**重复**的试验. 此时, 对应于给定处理与区组, 将有多观测值. 假设对应于每个处理与区组有 c 个观测值; 当重复个数不等时, 可做适当的调整. 由于出现了重复, 模型(35)必须做以下修正

$$X_{jkl} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \epsilon_{jkl} \quad (44)$$

X_{jkl} 的下标 j, k, l 分别对应于第 j 个处理, 第 k 个区组, 第 l 次重复. 在(44)式中, μ, α_j, β_k 如前所定义, ϵ_{jkl} 表示随机误差, γ_{jk} 表示行-列(或处理-区组)的**交互效应**, 通常简称为**交互**. 它们之间的限制如下

$$\sum_j \alpha_j = 0 \quad \sum_k \beta_k = 0 \quad \sum_j \gamma_{jk} = 0 \quad \sum_k \gamma_{jk} = 0 \quad (45)$$

并假定 X_{jkl} 服从均值为 $\mu + \alpha_j + \beta_k + \gamma_{jk}$, 方差为 σ^2 的正态分布.

如前所示, 总变差可以分解为处理间变差 V_R , 区组间变差 V_C , 交互变差 V_I 和残差 V_E :

$$V = V_R + V_C + V_I + V_E \quad (46)$$

其中

$$V = \sum_{j,k,l} (X_{jkl} - \bar{X})^2 \quad (47)$$

$$V_R = bc \sum_{j=1}^a (\bar{X}_{j..} - \bar{X})^2 \quad (48)$$

$$V_C = ac \sum_{k=1}^b (\bar{X}_{.k.} - \bar{X})^2 \quad (49)$$

$$V_I = c \sum_{j,k} (\bar{X}_{jk.} - \bar{X}_{j..} - \bar{X}_{.k.} + \bar{X})^2 \quad (50)$$

$$V_E = \sum_{j,k,l} (X_{jkl} - \bar{X}_{jk.})^2 \quad (51)$$

上述各式中, 下标的点表示对相应的下标已被求和. 例如

$$\bar{X}_{j..} = \frac{1}{bc} \sum_{k,l} X_{jkl} = \frac{1}{b} \sum_k \bar{X}_{jk.} \quad (52)$$

变差的数学期望也可如前计算. 若变差的自由度均已知, 则可建立表 16.6 的方差分析表, 表的最后一列可用来检验零假设:

$H_0^{(1)}$: 所有处理均值相等, 即 $\alpha_j = 0$.

$H_0^{(2)}$: 所有区组均值相等, 即 $\beta_k = 0$.

$H_0^{(3)}$: 处理与区组间无交互作用, 即 $\gamma_{jk} = 0$.

表 16.6

变 差	自由度	均 方	F
处理间变差 V_R	$a - 1$	$\hat{S}_R^2 = \frac{V_R}{a - 1}$	$\frac{\hat{S}_R^2}{\hat{S}_E^2} \sim F(a - 1, ab(c - 1))$
区组间变差 V_C	$b - 1$	$\hat{S}_C^2 = \frac{V_C}{b - 1}$	$\frac{\hat{S}_C^2}{\hat{S}_E^2} \sim F(b - 1, ab(c - 1))$
交互变差 V_I	$(a - 1)(b - 1)$	$\hat{S}_I^2 = \frac{V_I}{(a - 1)(b - 1)}$	$\frac{\hat{S}_I^2}{\hat{S}_E^2} \sim F((a - 1)(b - 1), ab(c - 1))$
残差 V_E	$ab(c - 1)$	$\hat{S}_E^2 = \frac{V_E}{ab(c - 1)}$	
总变差 V	$abc - 1$		

若知道具体观测值,在给定显著性水平下,应首先利用统计量 $\frac{\hat{S}_I^2}{\hat{S}_E^2}$ 判断是否拒绝 $H_0^{(3)}$. 此时可能存在两个问题:

1. 不能拒绝 $H_0^{(3)}$. 此时可认为交互作用不是很大. 因此,可以通过统计量 $\frac{\hat{S}_R^2}{\hat{S}_E^2}$ 和 $\frac{\hat{S}_C^2}{\hat{S}_E^2}$ 来检验 $H_0^{(1)}, H_0^{(2)}$. 有些统计学家认为此时可合并 V_I 和 V_E 得到均方

$$(V_I + V_E) / [(a - 1)(b - 1) + ab(c - 1)]$$

然后利用它来替换 F 检验中的 \hat{S}_E^2 .

2. 拒绝 $H_0^{(3)}$. 此时知交互作用非常显著. 仅当因素的差异比起交互作用来说很大时,才可认为因素的差异具有重要性. 因此,许多统计学家认为用 $\frac{\hat{S}_R^2}{\hat{S}_I^2}$ 和 $\frac{\hat{S}_C^2}{\hat{S}_I^2}$ 来检验 $H_0^{(1)}, H_0^{(2)}$ 比用表 16.6 提供的统计量要好些. 我们也采用这种统计量.

重复试验下的方差分析,可通过将同一个处理(行)与区组(列)下的重复观测值汇总成一个观测值,得到一个在各个处理与区组下都只有一个观测值的双因素方差分析表,如表 16.5. 由习题 16.13 可看到具体的操作过程.

实验设计

上述方差分析的方法是在试验的基础上获得的. 为了得到尽可能多的信息,试验设计必须提前认真进行,这就是我们经常提到的**实验设计**. 下面是实验设计的几个重要例子:

1. **完全随机化**. 假设我们有一个如例 1 所示的农业试验. 为了设计此试验,可把一块地分为 $4 \times 4 = 16$ 个小块(如图 16-1 所示的正方形,实际上可以是任意形状),每个处理(用 A, B, C, D 表示)分配到完全随机选择的 4 个区组中. 随机化的目的是为了减少各种误差源,例如土壤的肥沃度.

D	A	C	C
B	D	B	A
D	C	B	D
A	B	C	A

完全随机化

图 16-1

I	C	B	A	D
II	A	B	D	C
III	B	C	D	A
IV	A	D	C	B

随机化区组

图 16-2

2. **随机化区组**. 如同例 2 所示,对每一区组都必须有处理的完全集. 对于每一个区组(用 I、II、III、IV 表示),随机引进处理 A, B, C, D ,因此区组称为**随机化区组**. 这样设计的目的是为了控制**单源误差**,即区组间的差异.

3. **拉丁方**. 有时需同时控制**双源误差**,例如行间和列间差异. 例如在例 1 的试验中,不同行不同列间的误差可能是由于土壤不同处的肥沃度不同而造成的. 此时各行各列同一处理仅出

现一次显然较好,如图 16-3. 因为这种设计用到拉丁字母 A, B, C, D, 所以称为**拉丁方**.

D	B	C	A
B	D	A	C
C	A	D	B
A	C	B	D

拉丁方

图 16-3

B_γ	A_β	D_δ	C_α
A_δ	B_α	C_γ	D_β
D_α	C_δ	B_β	A_γ
C_β	D_γ	A_α	B_δ

希腊-拉丁方

图 16-4

4. **希腊-拉丁方**. 假如想控制三源误差, 就要用到图 16-4 的**希腊-拉丁方**, 它相当于两个拉丁方的叠加, 一个用拉丁字母 A, B, C, D, 另一个用希腊字母 $\alpha, \beta, \gamma, \delta$. 它要求每个拉丁字母和每个希腊字母只能同时出现一次. 这种正方形称为**正交的**.

习题及解答

单向分类或单因素试验

16.1 证明 $V = V_W + V_B$, 即,

$$\sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 + \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2$$

解 已知

$$X_{jk} - \bar{X} = (X_{jk} - \bar{X}_{j.}) + (\bar{X}_{j.} - \bar{X})$$

平方再对 j, k 求和, 则

$$\begin{aligned} \sum_{j,k} (X_{jk} - \bar{X})^2 &= \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 + \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 \\ &\quad + 2 \sum_{j,k} (X_{jk} - \bar{X}_{j.})(\bar{X}_{j.} - \bar{X}) \end{aligned}$$

下面证明上式最后一项为 0:

$$\begin{aligned} \sum_{j,k} (X_{jk} - \bar{X}_{j.})(\bar{X}_{j.} - \bar{X}) &= \sum_{j=1}^a (\bar{X}_{j.} - \bar{X}) \left[\sum_{k=1}^b (X_{jk} - \bar{X}_{j.}) \right] \\ &= \sum_{j=1}^a (\bar{X}_{j.} - \bar{X}) \left(\sum_{k=1}^b X_{jk} - b\bar{X}_{j.} \right) = 0 \end{aligned}$$

因为

$$\bar{X}_{j.} = \frac{1}{b} \sum_{k=1}^b X_{jk}$$

16.2 用本章(1), (2), (13)式的记号, 验证(a) $T = ab \bar{X}$, (b) $T_{j.} = b \bar{X}_{j.}$, (c) $\sum_j T_{j.} = ab \bar{X}$.

解 (a)

$$T = \sum_{j,k} X_{jk} = ab \left(\frac{1}{ab} \sum_{j,k} X_{jk} \right) = ab \bar{X}$$

(b)

$$T_{j.} = \sum_k X_{jk} = b \left(\frac{1}{b} \sum_k X_{jk} \right) = b \bar{X}_{j.}$$

(c) 因为 $T_{j.} = \sum_k X_{jk}$, 由(a)知

$$\sum_j T_{j.} = \sum_j \sum_k X_{jk} = T = ab \bar{X}$$

16.3 证明本章的快捷公式(10), (11)和(12).

证明

$$V = \sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk}^2 - 2\bar{X}X_{jk} + \bar{X}^2)$$

$$\begin{aligned}
&= \sum_{j,k} X_{jk}^2 - 2\bar{X} \sum_{j,k} X_{jk} + ab\bar{X}^2 = \sum_{j,k} X_{jk}^2 - 2\bar{X}(ab\bar{X}) + ab\bar{X}^2 \\
&= \sum_{j,k} X_{jk}^2 - ab\bar{X}^2 = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab}
\end{aligned}$$

同理

$$\begin{aligned}
V_B &= \sum_{j,k} (\bar{X}_{j\cdot} - \bar{X})^2 = \sum_{j,k} (\bar{X}_{j\cdot}^2 - 2\bar{X} \cdot \bar{X}_{j\cdot} + \bar{X}^2) \\
&= \sum_{j,k} \bar{X}_{j\cdot}^2 - 2\bar{X} \sum_{j,k} \bar{X}_{j\cdot} + ab\bar{X}^2 = \sum_{j,k} \left(\frac{T_{j\cdot}}{b} \right)^2 - 2\bar{X} \sum_{j,k} \left(\frac{T_{j\cdot}}{b} \right) + ab\bar{X}^2 \\
&= \frac{1}{b^2} \sum_{j=1}^a \sum_{k=1}^b T_{j\cdot}^2 - 2\bar{X}(ab\bar{X}) + ab\bar{X}^2 = \frac{1}{b} \sum_{j=1}^a T_{j\cdot}^2 - ab\bar{X}^2 \\
&= \frac{1}{b} \sum_{j=1}^a T_{j\cdot}^2 - \frac{T^2}{ab}
\end{aligned}$$

由 $V = V_W + V_B$, 或 $V_W = V - V_B$ 即得(12)式.

16.4 现有一经过化学药品 A, B, C 处理过的特定类型的土壤, 表 16.7 表示种植在其上的某品种小麦的亩产量(单位: 蒲式耳). 利用公式(7)和(8), 计算:

(a) 各个处理的均值, (b) 总均值, (c) 总变差, (d) 组间变差, (e) 组内变差, (f) 给出表 16.7 数据的 Minitab 分析, 并指出(a)到(e)的输出结果.

表 16.7

A	48	49	50	49
B	47	49	48	48
C	49	51	50	50

解 为了简化计算, 可以从表 16.7 的数据中均减去一个适当的数, 例如 45, 这样做并不影响变差值, 从而得到表 16.8.

表 16.8

3	4	5	4
2	4	3	3
4	6	5	5

(a) 表 16.8 的处理均值分别为

$$\bar{X}_{1\cdot} = \frac{1}{4}(3+4+5+4) = 4,$$

$$\bar{X}_{2\cdot} = \frac{1}{4}(2+4+3+3) = 3,$$

$$\bar{X}_{3\cdot} = \frac{1}{4}(4+6+5+5) = 5$$

因此, 通过加 45, 即可得 A, B, C 三种处理的平均产量, 分别为 49, 48, 50 蒲式耳/英亩.

(b) 总均值为

$$\bar{X} = \frac{1}{12}(3+4+5+4+2+4+3+3+4+6+5+5) = 4$$

因此表 16.7 所示数据的总均值为 $45+4=49$ 蒲式耳/英亩.

(c) 总变差为

$$\begin{aligned}
V &= \sum_{j,k} (X_{jk} - \bar{X})^2 = (3-4)^2 + (4-4)^2 + (5-4)^2 + (4-4)^2 + (2-4)^2 \\
&\quad + (4-4)^2 + (3-4)^2 + (3-4)^2 + (4-4)^2 + (6-4)^2 \\
&\quad + (5-4)^2 + (5-4)^2 = 14
\end{aligned}$$

(d) 组间变差为

$$V_B = b \sum_j (\bar{X}_{j.} - \bar{X})^2 = 4[(4-4)^2 + (3-4)^2 + (5-4)^2] = 8$$

(e) 组内变差为

$$V_W = V - V_B = 14 - 8 = 6$$

另解

$$\begin{aligned} V_W &= \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 = (3-4)^2 + (4-4)^2 + (5-4)^2 + (4-4)^2 + (2-3)^2 \\ &\quad + (4-3)^2 + (3-3)^2 + (3-3)^2 + (4-5)^2 + (6-5)^2 \\ &\quad + (5-5)^2 + (5-5)^2 = 6 \end{aligned}$$

注 表 16.9 是习题 16.4, 16.5, 16.6 的方差分析表.

表 16.9

变 差	自由度	均 方	F
组间变差 $V_B = 8$	$a - 1 = 2$	$\hat{S}_B^2 = \frac{8}{2} = 4$	$\frac{\hat{S}_B^2}{\hat{S}_W^2} = \frac{4}{2/3} = 6$ 服从自由度为 2, 9 的 F 分布
组内变差 $V_W = V - V_B = 14 - 8 = 6$	$a(b - 1) = 3 \times 3 = 9$	$\hat{S}_W^2 = \frac{6}{9} = \frac{2}{3}$	
总变差 $V = 14$	$ab - 1 = 3 \times 4 - 1 = 11$		

(f) Minitab 计算结果如下. 表 16.7 的数据输入到工作表的 3 列中, 各列分别记为 A, B, C.

MTB>AOVOneWay 'A' 'B' 'C'.

One-way Analysis of Variance

Analysis of Variance

Source	DF	SS	MS	F	P
Factor	2	8.000	4.000	6.00	0.022
Error	9	6.000	0.667		
Total	11	14.000			

Individual 95% CIs For Mean

Based on Pooled StDev

Level	N	Mean	StDev	
A	4	49.000	0.816	(- - - - - * - - - - -)
B	4	48.000	0.816	(- - - - - * - - - - -)
C	4	50.000	0.816	(- - - - - * - - - - -)

Pooled StDev = 0.816 48.0 49.2 50.4

从 Minitab 输出可看出, 3 个处理均值分别为: 49, 48 及 50, 与 (a) 中的结论一致; (c) 的总变差 $V = 14$, 对应于 Minitab 输出中单因素方差分析表的 Total SS = 14; (d) 的结论 $V_B = 8$, 对应于 Factor SS = 8; (e) 的结论 $V_W = 6$ 对应于 Error SS = 6. Minitab 也提供了数据的点图与箱形图, 见图 16.5、图 16.6.

16.5 对于习题 16.4, 试寻找一个总体方差 σ^2 的无偏估计.

(a) 在零假设成立时 (即各处理均值相等的条件下), 利用组间变差.

(b) 利用组内变差.

(c) 参考习题 16.4 解的 Minitab 输出, 指出对应于 (a), (b) 的方差估计.

解 (a)

$$\hat{S}_B^2 = \frac{V_B}{a - 1} = \frac{8}{3 - 1} = 4$$

(b)

$$\hat{S}_w^2 = \frac{V_w}{a(b-1)} = \frac{6}{3(4-1)} = \frac{2}{3}$$

(c) 方差估计 \hat{S}_B^2 和 Minitab 输出中因素均方一致, 即 Factor MS=4.000; 方差估计 \hat{S}_w^2 和 Minitab 输出中误差均方一致, 即 Error MS=0.667.

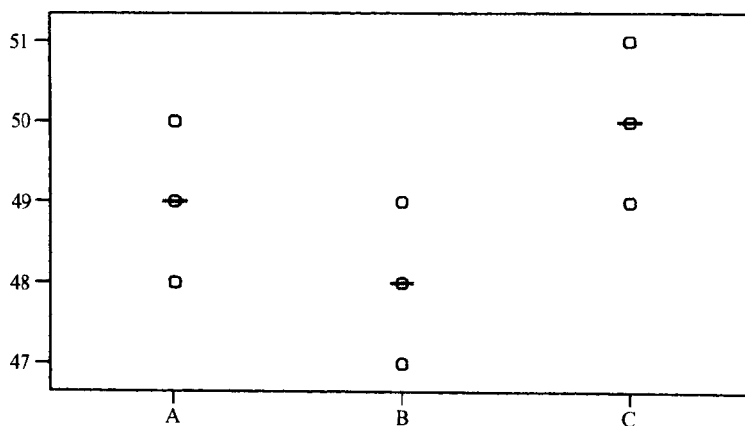


图 16-5 A-C 点图(组均值用直线表示).

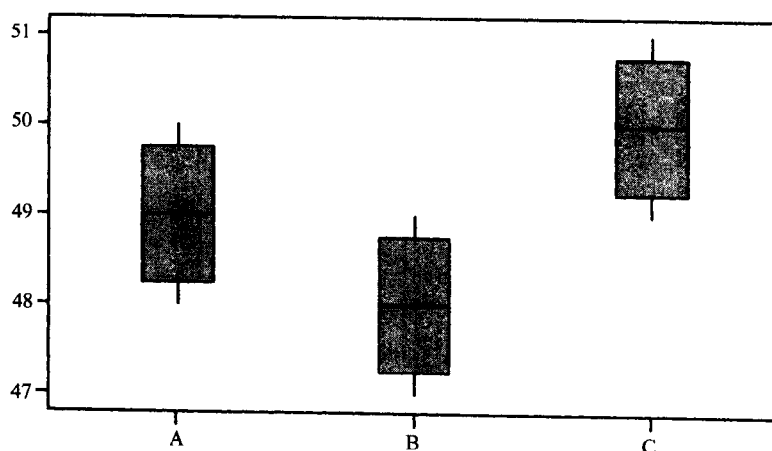


图 16-6 A-C 箱形图(均值用实心点表示).

- 16.6 见习题 16.4, 看是否能在如下的显著性水平(a) 0.05, (b) 0.01 下, 拒绝原假设(即处理均值相等)? (c) 参考习题 16.4 解的 Minitab 输出, 检验零假设是否成立?

解 已知

$$F = \frac{\hat{S}_B^2}{\hat{S}_w^2} = \frac{4}{2/3} = 6$$

自由度为 $a-1=3-1=2$, 及 $a(b-1)=3(4-1)=9$.

(a) 由附录 V, 可查得 $F_{0.95}(2, 9)=4.26$, 而 $F=6 > F_{0.95}$, 因此可在水平 0.05 下, 拒绝零假设, 即处理间均值不等.

(b) 由附录 VI, 可查得 $F_{0.99}(2, 9)=8.02$, 而 $F=6 < F_{0.99}$, 因此可在水平 0.01 下, 不能拒绝零假设, 即处理间均值相等.

(c) 由习题 16.4 解的 Minitab 输出, 可知 F 值为 6, p -值为 0.022, 即拒绝零假设的最小显著性水平应为 0.022, 因此当显著性水平为 0.05 时, 拒绝零假设; 但水平为 0.01 时, 不能拒绝零假设.

- 16.7 利用公式(10), (11)和(12)求解习题 16.4.

解 把表 16.8 的数据重新安排即可得到表 16.10.

解 为了简化计算,可以从表 16.11 的数据中均减去一个适当的数,例如 60,从而得到表 16.12.

$$V = 2658 - \frac{54^2}{5 \times 5} = 2658 - 116.64 = 2541.36$$

$$V_B = \frac{3874}{5} - \frac{54^2}{5 \times 5} = 774.8 - 116.64 = 658.16$$

由此得到表 16.13,查表知 $F_{0.95}(4,20) = 2.87$,因此不能在水平 0.05 下拒绝零假设,当然水平 0.01 下更不能拒绝零假设.

表 16.13

变 差	自由度	均 方	F
组间变差 $V_B = 658.2$	$a - 1 = 4$	$\hat{S}_B^2 = \frac{658.2}{4} = 164.5$	$F = \frac{164.5}{94.16} = 1.75$
组内变差 $V_W = 1883.2$	$a(b - 1) = 5 \times 4 = 20$	$\hat{S}_W^2 = \frac{1883.2}{20} = 94.16$	
总变差 $V = 2514.4$	$ab - 1 = 24$		

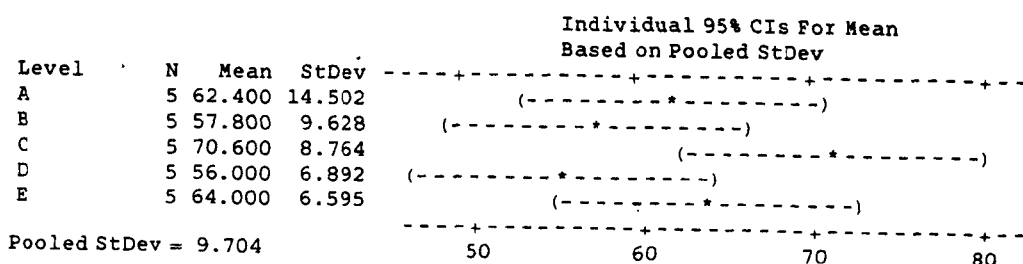
把每台机器的产量作为一列输入到 Minitab 工作表中,共有 5 列,用命令 AOVOneway 'A' 'B' 'C' 'D' 'E',可得到如下输出:

MTB>AOVOneway 'A' 'B' 'C' 'D' 'E'

One-way Analysis of Variance

Analysis of Variance

Source	DF	SS	MS	F	P
Factor	4	658.0	164.5	1.75	0.179
Error	20	1883.2	94.2		
Total	24	2541.4			



p-值为 0.179,即零假设被拒绝的最小显著性水平为 0.179.因此,在水平 0.01 或 0.05 下,零假设均不会被拒绝.

观测值数目不等时所作的修正

16.9 表 16.14 是某公司生产的三种型号的电子管的寿命(单位:小时).用(3),(8)式的长方方法,在显著性水平(a)0.05,(b)0.01 下判断三种电子管之间是否存在差异?

表 16.14

样本 1	407	411	409		
样本 2	404	406	408	405	402
样本 3	410	408	406	408	

解 从表 16.14 的数据中均减去 400,得到表 16.15,此表列出了行总和,样本均值,总均值.

$$V = \sum_{j,k} (X_{jk} - \bar{X})^2 = (7-7)^2 + (11-7)^2 + \cdots + (8-7)^2 = 72$$

$$V_B = \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 = \sum_j N_j (\bar{X}_{j.} - \bar{X})^2 \\ = 3(9-7)^2 + 5(7-5)^2 + 4(8-7)^2 = 36$$

$$V_W = V - V_B = 72 - 36 = 36$$

也可直接计算 V_W

$$V_W = (7-9)^2 + (11-9)^2 + (9-9)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2 \\ + (5-5)^2 + (2-5)^2 + (10-8)^2 + (8-8)^2 + (6-8)^2 + (8-8)^2 = 36$$

表 16.15

						总和	均值
样本 1	7	11	9			27	9
样本 2	4	6	8	5	2	25	5
样本 3	10	8	6	8		32	8
$\bar{X} = \text{总均值} = \frac{84}{12} = 7$							

表 16.16 为上述数据的方差分析表. 由附录 V 可查得 $F_{0.95}(2, 9) = 4.26$, 由附录 VI 可查得 $F_{0.99}(2, 9) = 8.02$, 因此可在水平 0.05 下拒绝零假设; 样本间均值相等 (即认为三种电子管间有差异), 但当水平为 0.01 时, 不能拒绝零假设 (即认为三种电子管间无差异).

表 16.16

变差	自由度	均方	F
$V_B = 36$	$a - 1 = 2$	$\hat{S}_B^2 = \frac{36}{2} = 18$	$\frac{\hat{S}_B^2}{\hat{S}_W^2} = \frac{18}{4} = 4.5$
$V_W = 36$	$N - a = 9$	$\hat{S}_W^2 = \frac{36}{9} = 4$	

16.10 用快捷方法(24), (25), (26)式求解习题 16.9.

解 从表 16.15, 可知 $N_1 = 3, N_2 = 5, N_3 = 4, N = 12, T_{1.} = 27, T_{2.} = 25, T_{3.} = 32, T = 84$, 因此

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{N} = 7^2 + 11^2 + \cdots + 6^2 + 8^2 - \frac{84^2}{12} = 72$$

$$V_B = \sum_j \frac{T_{j.}^2}{N_j} - \frac{T^2}{N} = \frac{27^2}{3} + \frac{25^2}{5} + \frac{32^2}{4} - \frac{84^2}{12} = 36$$

$$V_W = V - V_B = 36$$

方差分析过程与习题 16.9 一样.

双因素方差分析

16.11 表 16.17 表示种植在三块用不同肥料处理过的土壤中的四种不同品种庄稼的亩产量. 用长方法, 在显著性水平 0.01 下判断亩产量是否 (a) 由于施肥的不同, (b) 由于庄稼品种的不同而存在差异? (c) 用 Minitab 求解.

表 16.17

	庄稼 I	庄稼 II	庄稼 III	庄稼 IV
肥料 A	4.5	6.4	7.2	6.7
肥料 B	8.8	7.8	9.6	7.0
肥料 C	5.9	6.8	5.7	5.2

解 计算行之和, 行均值, 列之和, 列均值, 数据总和, 总均值, 见表 16.18.

表 16.18

	庄稼 I	庄稼 II	庄稼 III	庄稼 IV	行之和	行均值
肥料 A	4.5	6.4	7.2	6.7	24.8	6.2
肥料 B	8.8	7.8	9.6	7.0	33.2	8.3
肥料 C	5.9	6.8	5.7	5.2	23.6	5.9
列之和	19.2	21.0	22.5	18.9	总和 = 81.6	
列均值	6.4	7.0	7.5	6.3	总均值 = 6.8	

行均值与总均值间的变差为

$$V_R = 4[(6.2 - 6.8)^2 + (8.3 - 6.8)^2 + (5.9 - 6.8)^2] = 13.68$$

列均值与总均值间的变差为

$$V_C = 3[(6.4 - 6.8)^2 + (7.0 - 6.8)^2 + (7.5 - 6.8)^2 + (6.3 - 6.8)^2] = 2.82$$

总变差为

$$\begin{aligned} V &= (4.5 - 6.8)^2 + (6.4 - 6.8)^2 + (7.2 - 6.8)^2 + (6.7 - 6.8)^2 \\ &\quad + (8.8 - 6.8)^2 + (7.8 - 6.8)^2 + (9.6 - 6.8)^2 + (7.0 - 6.8)^2 + (5.9 - 6.8)^2 \\ &\quad + (6.8 - 6.8)^2 + (5.7 - 6.8)^2 + (5.2 - 6.8)^2 = 23.08 \end{aligned}$$

随机变差

$$V_E = V - V_R - V_C = 6.58$$

表 16.19 为方差分析表.

表 16.19

变差	自由度	均方	F
$V_R = 13.68$	2	$S_R^2 = 6.84$	$\frac{S_R^2}{S_E^2} = 6.24$, 自由度为 2, 6
$V_C = 2.82$	3	$S_C^2 = 0.94$	$\frac{S_C^2}{S_E^2} = 0.86$, 自由度为 3, 6
$V_E = 6.58$	6	$S_E^2 = 1.097$	
$V = 23.08$	11		

显著性水平为 0.05 时, 查表知 $F_{0.95}(2, 6) = 5.14$. 由于 $5.14 < 6.24$, 则可拒绝行均值相等的假设, 认为在水平 0.05 下, 由于施肥的不同产量间存在着显著差异.

由于列均值所对应的 F 值小于 1, 可认为庄稼品种的不同不会影响亩产量.

(c) 首先给出 Minitab 工作表的数据结构, 然后给出双因素的 Minitab 分析.

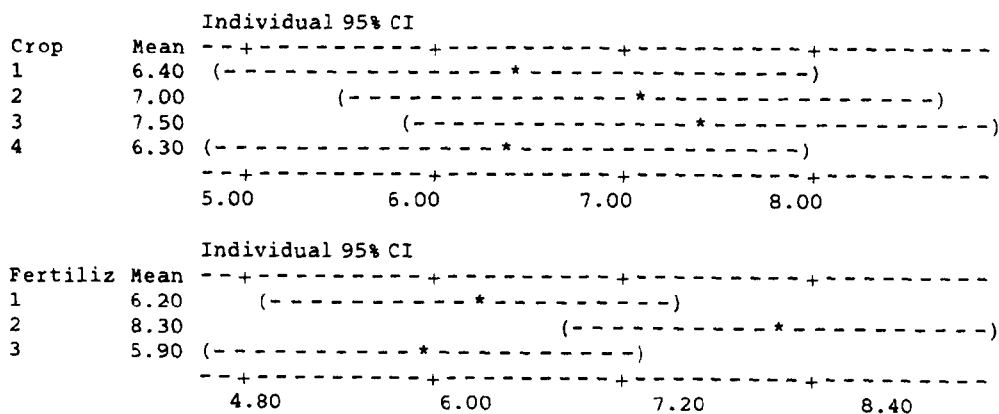
Row	Crop	Fertilizer	Yield
1	1	1	4.5
2	1	2	8.8
3	1	3	5.9
4	2	1	6.4
5	2	2	7.8
6	2	3	6.8
7	3	1	7.2
8	3	2	9.6
9	3	3	5.7
10	4	1	6.7
11	4	2	7.0
12	4	3	5.2

```
MTB>Twoway 'Yield' 'Crop' 'Fertilizer';
SUBC>Means 'Crop' 'Fertilizer'.
```

Two-way Analysis of Variance

Analysis of Variance For Yield

Source	DF	SS	MS	F	P
Crop	3	2.82	0.94	0.86	0.512
Fertiliz	2	13.68	6.84	6.24	0.034
Error	6	6.58	1.10		
Total	11	23.08			



工作表中的数据必须与表 16.17 中的数据保持一致. 第一行 1, 1, 4.5, 对应于庄稼 I, 肥料 1 下的产量 4.5; 第二行 1, 2, 8.8, 对应于庄稼 I, 肥料 2 下的产量 8.8. 以下同理. 运用统计软件最常犯的一个错误是误建数据结构, 因此一定要保证工作表中的数据和表 16.17 中的数据一致. 我们还注意到 Minitab 的双因素方差分析包含了表 16.19 相同的信息. Minitab 输出给出的 p -值允许观测者不用查 F 分布表, 就可对假设进行检验. 庄稼所对应的 p -值为 0.512, 它是由于庄稼品种不同因而拒绝产量间存在差异的最小显著性水平, 4 品种庄稼的产量在水平 0.05, 0.01 下都没有显著的差异. 肥料因素所对应的 p -值为 0.034, 从而我们得到结论: 三种施肥下的亩产量在水平 0.05 下有显著差异, 在水平 0.01 下差异不显著.

从 Minitab 输出中四种庄稼产量的置信区间可以断定: 庄稼品种不同并不影响产量, 而三种肥料的置信区间又表明, 肥料 B 作用下的产量可能要高于肥料 A, C 下的产量.

16.12 用快捷公式计算习题 16.11.

解 从表 16.18 可知

$$\begin{aligned}\sum_{j,k} X_{jk}^2 &= 4.5^2 + 6.4^2 + \cdots + 5.2^2 = 577.96 \\ T &= 24.8 + 33.2 + 23.6 = 81.6 \\ \sum_j T_j^2 &= 24.8^2 + 33.2^2 + 23.6^2 = 2274.24 \\ \sum_k T_k^2 &= 19.2^2 + 21.0^2 + 22.5^2 + 18.9^2 = 1673.10\end{aligned}$$

因此

$$\begin{aligned}V &= \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} = 577.96 - 554.88 = 23.08 \\ V_R &= \frac{1}{b} \sum_j T_j^2 - \frac{T^2}{ab} = \frac{1}{4} \times 2274.24 - 554.88 = 13.68 \\ V_C &= \frac{1}{a} \sum_k T_k^2 - \frac{T^2}{ab} = \frac{1}{3} \times 1673.10 - 554.88 = 2.82 \\ V_E &= V - V_R - V_C = 23.08 - 13.68 - 2.82 = 6.58\end{aligned}$$

和习题 16.11 得到结论一致.

有重复的双因素试验

16.13 某厂商希望了解四种生产螺栓的机器(A, B, C, D)的效率, 为此设计了如下试验: 在

一周内,记录两个班组每天每类机器生产的次品数.结论见表 16.20.利用方差分析的方法,在显著性水平 0.05 下,判断(a) 机器间是否存在差异? (b) 班组间是否存在差异? (c) 用 Minitab 进行方差分析,并用 p -值求解(a)与(b).

表 16.20

机 器	班组 I					班组 II				
	星期一	星期二	星期三	星期四	星期五	星期一	星期二	星期三	星期四	星期五
A	6	4	5	5	4	5	7	4	6	8
B	10	8	7	7	9	7	9	12	8	8
C	7	5	6	5	9	9	7	5	4	6
D	8	4	6	5	5	5	7	9	7	10

解 数据可整理成表 16.21,此表中包含有两个主因子机器与班组.因为有两个班组,所以可以认为一周内的每天都产生了一次重复.因此表 16.21 数据的总变差为

$$V = 6^2 + 4^2 + 5^2 + \cdots + 7^2 + 10^2 - \frac{268^2}{40} = 1946 - 1795.6 = 150.4$$

表 16.21

因素 I: 机器	因素 II: 班组	重复					和
		星期一	星期二	星期三	星期四	星期五	
A	1	6	4	5	5	4	24
	2	5	7	4	6	8	30
B	1	10	8	7	7	9	41
	2	7	9	12	8	8	44
C	1	7	5	6	5	9	32
	2	9	7	5	4	6	31
D	1	8	4	6	5	5	28
	2	5	7	9	7	10	38
和		57	51	54	47	59	268

为了对两个主因子(机器与班组)进行分析,仅考虑对应于各因子从星期一到星期五的数据和.例如机器 A、班组 I 所在行之和为 24,其他数据见表 16.22,它是一个无重复试验数据的双因子表.表 16.22 的总变差称为子总变差,用 V_s 表示

表 16.22

机器	班组 I	班组 II	和
A	24	30	54
B	41	44	85
C	32	31	63
D	28	38	66
和	125	143	268

$$V_s = \frac{24^2}{5} + \frac{41^2}{5} + \frac{32^2}{5} + \frac{28^2}{5} + \frac{30^2}{5} + \frac{44^2}{5} + \frac{31^2}{5} + \frac{38^2}{5} - \frac{268^2}{40}$$

$$= 1861.2 - 1795.6 = 65.6$$

行间变差为

$$V_R = \frac{54^2}{10} + \frac{85^2}{10} + \frac{63^2}{10} + \frac{66^2}{10} - \frac{268^2}{40} = 1846.6 - 1795.6 = 51.0$$

列间变差为

$$V_C = \frac{125^2}{20} + \frac{143^2}{20} - \frac{268^2}{40} = 1803.7 - 1795.6 = 8.1$$

假如从子总变差 V_S 中减去行及列间变差 $V_R + V_C$, 就可得到由于行列的交互作用而产生的变差:

$$V_I = V_S - V_R - V_C = 65.6 - 51.0 - 8.1 = 6.5$$

最后考虑残差, 即随机变差(假设一周各天之间并无差异), 可以通过总变差 V 减去子总变差得到, 即

$$V_E = V - (V_R + V_C + V_I) = V - V_S = 150.4 - 65.6 = 84.8$$

上述所求方差见表 16.23 的方差分析表. 此表也给出每种变差的自由度. 因为表 16.22 有 4 行, 所以行间变差的自由度为 $4 - 1 = 3$; 而两列的列间变差的自由度为 $2 - 1 = 1$. 为了得到交互变差的自由度, 首先考虑表 16.22 的输入数据的个数, 此例中为 8, 所以子总变差的自由度为 $8 - 1 = 7$, 其中包含有行自由度 3 及列自由度 1, 剩余的 $(7 - (3 + 1) = 3)$ 即为交互变差的自由度. 原始数据表 16.21 有 40 个输入数据, 则总自由度为 $40 - 1 = 39$. 因此, 残差的自由度为 $39 - 7 = 32$.

表 16.23

变差	自由度	均方	F
行间变差(机器) $V_R = 51.0$	3	$\hat{S}_R^2 = 17.0$	$\frac{17.0}{2.65} = 6.42$
列间变差(班组) $V_C = 8.1$	1	$\hat{S}_C^2 = 8.1$	$\frac{8.1}{2.65} = 3.06$
交互变差 $V_I = 6.5$	3	$\hat{S}_I^2 = 2.167$	$\frac{2.167}{2.65} = 0.817$
子总变差 $V_S = 65.6$	7		
残差 $V_E = 84.8$	32	$\hat{S}_E^2 = 2.65$	
总变差 $V = 150.4$	39		

首先判断交互作用是否显著. 查表并通过内插知临界值 $F_{0.95}(3, 32) = 2.90$, 而计算的 F 值为 0.817, 故交互作用不显著. 机器间存在显著差异, 因为计算的 F 值为 6.42 而临界值 $F_{0.95}(3, 32) = 2.90$. 班组的临界值为 4.15, 而计算的 F 值为 3.06, 因而不会由于班组的不同而引起次品上的差异.

Minitab 的数据结构如下, 把它和表 16.21 的数据进行比较, 看看他们是如何匹配的?

Row	Machine	Shift	Defects
1	1	1	6
2	1	1	4
3	1	1	5
4	1	1	5
5	1	1	4
6	1	2	5
7	1	2	7
8	1	2	4
9	1	2	6
10	1	2	8
11	2	1	10
12	2	1	8
13	2	1	7
14	2	1	7
15	2	1	9
16	2	2	7

17	2	2	9
18	2	2	12
19	2	2	8
20	2	2	8
21	3	1	7
22	3	1	5
23	3	1	6
24	3	1	5
25	3	1	9
26	3	2	9
27	3	2	7
28	3	2	5
29	3	2	4
30	3	2	6
31	4	1	8
32	4	1	4
33	4	1	6
34	4	1	5
35	4	1	5
36	4	2	5
37	4	2	7
38	4	2	9
39	4	2	7
40	4	2	10

命令 `MTB>Twoway 'Defects' 'Machine' 'shift'` 可用来进行双因素分析. 交互作用的 p -值为 0.494, 这是零假设被拒绝的最小显著性水平. 显然在班组和机器间交互作用不大. 班组的 p -值为 0.090, 大于 0.050, 即认为两个班组间的差异不显著; 机器的 p -值为 0.002, 在显著性水平 0.05 下, 由四种机器的不同而引起的次品数有显著差异.

`MTB>Twoway 'Defects' 'Machine' 'shifts';`

Two-way Analysis of Variance

Analysis of Variance for Variance

Source	DF	SS	MS	F	P
Machine	3	51.00	17.00	6.42	0.002
Shift	1	8.10	8.10	3.06	0.090
Interaction	3	6.50	2.17	0.82	0.494
Error	32	84.80	2.65		
Total	39	150.40			

图 16-7 是班组和机器的交互作用图. 此图指出机器和班组间可能存在的交互作用. 但是方差

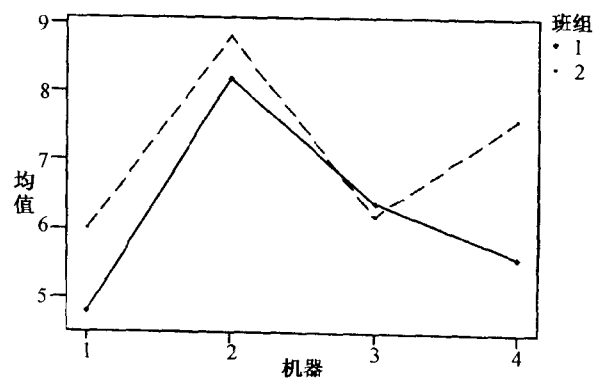


图 16-7 交互作用图——次品数均值

分析表却告诉我们二者之间交互作用不显著.当交互作用不存在时,班组 I 和班组 II 的折线图应该是平行的.图 16-8 的主效应图表明机器 1 在试验中平均次品数最少,机器 2 最多;班组 II 中的平均次品数比班组 I 要多.但方差分析表明班组间的差异并不显著.

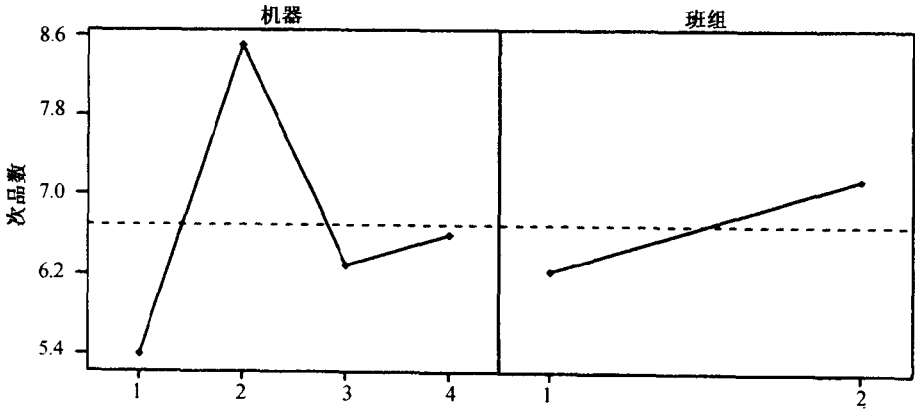


图 16-8 主效应图——次品数均值

16.14 在水平 0.01 下求解习题 16.13.

解 在此水平下,交互作用仍不显著,因此可进一步研究.

因为 $F_{0.99}(3, 32) = 4.47 < F = 6.42$ (行间),可认为在水平 0.01 下,各机器的效率不相等.

因为 $F_{0.99}(1, 32) = 7.51 > F = 3.06$ (列间),可认为在水平 0.01 下,班组间无显著差异.

拉丁方

16.15 某农民想检验四种不同肥料(A, B, C, D)对小麦产量的影响.为了消除土壤肥沃度变异之误差源,他设计了一个拉丁方,见表 16.24.其中的数字表示每单位面积的产量(蒲式耳).应用方差分析,在显著性水平(a) 0.05, (b) 0.01 下判断肥料间是否存在差异? (c) 给出此拉丁方的 Minitab 解.

表 16.24

A 18	C 21	D 25	B 11
D 22	B 12	A 15	C 19
B 15	A 20	C 23	D 24
C 22	D 21	B 10	A 17

表 16.25

				和
A 18	C 21	D 25	B 11	75
D 22	B 12	A 15	C 19	68
B 15	A 20	C 23	D 24	82
C 22	D 21	B 10	A 17	70
和	77	74	73	71
				295

表 16.26

	A	B	C	D	
和	70	48	85	92	295

解 首先计算各行各列之和,见表 16.25.每种肥料作用下的总产量见表 16.26.总变差、行间变差、列间变差及处理间(即肥料间)变差计算方法如前.则总变差为

$$V = 18^2 + 21^2 + 25^2 + \cdots + 10^2 + 17^2 - \frac{295^2}{16} = 5769 - 5439.06 = 329.94$$

行间变差为

$$V_R = \frac{75^2}{4} + \frac{68^2}{4} + \frac{82^2}{4} + \frac{70^2}{4} - \frac{295^2}{16} = 5468.25 - 5439.06 = 29.19$$

列间变差为

$$V_C = \frac{77^2}{4} + \frac{74^2}{4} + \frac{73^2}{4} + \frac{71^2}{4} - \frac{295^2}{16} = 5443.75 - 5439.06 = 4.69$$

处理间变差为

$$V_B = \frac{70^2}{4} + \frac{48^2}{4} + \frac{85^2}{4} + \frac{92^2}{4} - \frac{295^2}{16} = 5723.25 - 5439.06 = 284.19$$

表 16.27 为方差分析表.

表 16.27

变差	自由度	均方	F
行, 29.19	3	9.73	4.92
列, 4.69	3	1.563	0.79
处理, 284.19	3	94.73	47.9
残差, 11.87	6	1.978	
总变差, 329.94	15		

(a) 因为 $F_{0.95}(3, 6) = 4.76$, 可在水平 0.05 下, 拒绝行均值相等的假设, 即认为在显著性水平 0.05 下, 各行土壤的肥沃度有所差别.

因为列的 F 值小于 1, 故认为各列土壤的肥沃度没有差别.

因为处理的 F 值 $47.9 > 4.76$, 故认为肥料间存在差异.

(b) 因为 $F_{0.99}(3, 6) = 9.78$, 故在水平 0.01 下, 认为各行(或各列)土壤的肥沃度之间没有差异, 肥料间存在差异.

(c) 首先给出 Minitab 工作表的数据结构文件.

Row	Rows	Columns	Treatment	Yield
1	1	1	1	18
2	1	2	3	21
3	1	3	4	25
4	1	4	2	11
5	2	1	4	22
6	2	2	2	12
7	2	3	1	15
8	2	4	3	19
9	3	1	2	15
10	3	2	1	20
11	3	3	3	23
12	3	4	4	24
13	4	1	3	22
14	4	2	4	21
15	4	3	2	10
16	4	4	1	17

注意, 在输出中行与列均用 1, 2, 3, 4 表示, 表 16.24 的肥料 A, B, C, D 也分别用 1, 2, 3, 4 表示. 由命令 GLM 'Yield' = Rows Columns Treatment; 即可得到 Minitab 分析.

MTB> GLM 'Yield' = Rows Columns Treatment;

SUBC> SSquares 1;

SUBC> Brief 2.

General Linear Model

Factor	Type	Levels	Values			
		4	1	2	3	4
Rows	fixed	4	1	2	3	4
Columns	fixed	4	1	2	3	4
Treatment	fixed	4	1	2	3	4

Analysis of Variance for Yield, using Sequential SS for Tests

Source	DF	Seq SS	Adj SS	Seq MS	F	P
Rows	3	29.187	29.187	9.729	4.92	0.047
Columns	3	4.688	4.687	1.563	0.79	0.542
Treatment	3	284.187	284.187	94.729	47.86	0.000
Error	6	11.875	11.875	1.979		
Total	15	329.937				

Minitab 输出中 Seq MS 列和表 16.27 中均方 MS 列一致. 计算的 F 值和表 16.27 的 F 值也相同. 行、列及处理的 p -值分别为 0.047, 0.542, 0.000. 由 p -值的性质可知, 在水平 0.05 下, 行的均值间存在差异, 但在水平 0.01 下, 无差异; 列的均值在两种水平下都不存在差异; 在任何水平下, 肥料间都存在差异. 对四种肥料均值的进一步研究将会看到均值的差异到底有多大.

希腊-拉丁方

16.16 研究四种汽油 A, B, C, D 每加仑的行驶里程是否有区别, 这是件很有意义的事. 设计一个试验, 其中涉及到 4 个不同的司机, 4 辆不同的汽车, 4 条不同的马路.

解 因为汽油种数, 司机人数, 汽车数及马路的条数均相等, 所以可用希腊-拉丁方. 假设用行来表示汽车而列表示司机, 如表 16.28 所示. 现随机分派汽油 (A, B, C, D) 到各行各列, 但一定要使各个字母在每行每列只出现一次, 这样就可保证每个司机都有机会驾驶各辆汽车, 使用各种汽油. 而且每辆汽车不会使用同一种汽油.

现随机分派四条马路, 用 $\alpha, \beta, \gamma, \delta$ 表示. 按照拉丁方的方法来分派. 这样, 每个司机均有机会在四条马路上行驶. 表 16.28 表示了这种设计安排.

表 16.28

	司 机			
	1	2	3	4
汽车 1	B_γ	A_β	D_δ	C_α
汽车 2	A_δ	B_α	C_γ	D_β
汽车 3	D_α	C_δ	B_β	A_γ
汽车 4	C_β	D_γ	A_α	B_δ

16.17 假定在习题 16.16 中, 每加仑汽油的行驶里程见表 16.29, 在显著性水平 0.05 下, 利用方差分析的方法判断是否存在差异? 再用 Minitab 给出的 p -值进行分析.

表 16.29

	司 机			
	1	2	3	4
汽车 1	B_γ 19	A_β 16	D_δ 16	C_α 14
汽车 2	A_δ 15	B_α 18	C_γ 11	D_β 15
汽车 3	D_α 14	C_δ 11	B_β 21	A_γ 16
汽车 4	C_β 16	D_γ 16	A_α 15	B_δ 23

解 首先计算各行各列之和, 见表 16.30. 然后计算每一个拉丁字母和每一个希腊字母对应的数据之和, 即

$$A\text{ 之和: }15 + 16 + 15 + 16 = 62$$

$$B\text{ 之和: }19 + 18 + 21 + 23 = 81$$

$$C\text{ 之和: }16 + 11 + 11 + 14 = 52$$

$$D\text{ 之和: }14 + 16 + 16 + 15 = 61$$

$$\alpha\text{ 之和: }14 + 18 + 15 + 14 = 61$$

$$\beta\text{ 之和: }16 + 16 + 21 + 15 = 68$$

$$\gamma\text{ 之和: }19 + 16 + 11 + 16 = 62$$

$$\delta\text{ 之和: }15 + 11 + 16 + 23 = 65$$

再用快捷方法计算相应的变差
行

$$\frac{65^2}{4} + \frac{59^2}{4} + \frac{62^2}{4} + \frac{70^2}{4} - \frac{256^2}{16} = 4112.50 - 4096 = 16.50$$

列

$$\frac{64^2}{4} + \frac{61^2}{4} + \frac{63^2}{4} + \frac{68^2}{4} - \frac{256^2}{16} = 4102.50 - 4096 = 6.50$$

汽油(A, B, C, D)

$$\frac{62^2}{4} + \frac{81^2}{4} + \frac{52^2}{4} + \frac{61^2}{4} - \frac{256^2}{16} = 4207.50 - 4096 = 111.50$$

马路($\alpha, \beta, \gamma, \delta$)

$$\frac{61^2}{4} + \frac{68^2}{4} + \frac{62^2}{4} + \frac{65^2}{4} - \frac{256^2}{16} = 4103.50 - 4096 = 7.50$$

总变差为

$$V = 19^2 + 16^2 + 16^2 + \cdots + 15^2 + 23^2 - \frac{256^2}{16} = 4244 - 4096 = 148.00$$

随机变差为

$$148.00 - 16.50 - 6.50 - 111.50 - 7.50 = 6.00$$

结论见方差分析表 16.31. $N \times N$ 拉丁方的自由度为 $N^2 - 1$, 各行各列及拉丁、希腊字母的自由度为 $N - 1$, 因此误差的自由度为 $N^2 - 1 - 4(N - 1) = (N - 1)(N - 3)$. 本题中 $N = 4$.

表 16.30

					和
	$B\gamma$ 19	$A\beta$ 16	$D\delta$ 16	$C\alpha$ 14	65
	$A\delta$ 15	$B\alpha$ 18	$C\gamma$ 11	$D\beta$ 15	59
	$D\alpha$ 14	$C\delta$ 11	$B\beta$ 21	$A\gamma$ 16	62
	$C\beta$ 16	$D\gamma$ 16	$A\alpha$ 15	$B\delta$ 23	70
和	64	61	63	68	256

表 16.31

变差	自由度	均方	F
行(汽车), 16.50	3	5.500	$\frac{5.500}{2.000} = 2.75$
列(司机), 6.50	3	2.167	$\frac{2.167}{2.000} = 1.08$
汽油(A, B, C, D), 111.50	3	37.167	$\frac{37.167}{2.000} = 18.6$
马路($\alpha, \beta, \gamma, \delta$), 7.50	3	2.500	$\frac{2.500}{2.000} = 1.25$
残差, 6.00	3	2.000	
总变差, 148.00	15		

因为 $F_{0.95}(3,3) = 9.28$, $F_{0.99}(3,3) = 29.5$, 因此可认为在水平 0.05 下, 汽油间存在差异, 而水平为 0.01 时, 不存在差异.

下面是 Minitab 工作表的数据结构:

Row	Car	Driver	Casoline	Road	MPG
1	1	1	2	3	19
2	1	2	1	2	16
3	1	3	4	4	16
4	1	4	3	1	14
5	2	1	1	4	15
6	2	2	2	1	18
7	2	3	3	3	11
8	2	4	4	2	15
9	3	1	4	1	14
10	3	2	3	4	11
11	3	3	2	2	21
12	3	4	1	3	16
13	4	1	3	2	16
14	4	2	4	3	16
15	4	3	1	1	15
16	4	4	2	4	23

注意上述工作表中汽车及司机的表示与表 16.29 一样, 汽油 A, B, C, D 在工作表中用 1, 2, 3, 4 表示. 表 16.29 中的马路 $\alpha, \beta, \gamma, \delta$ 在工作表中用 1, 2, 3, 4 表示. 通过命令 MTB>GLM'MPG' = Car Driver Gasoline Road; 即可进行 Minitab 分析.

MTB>GLM 'MPG' = Car Driver Gasoline Road;

SUBC>SSquares 1;

SUBC>Brief 2.

General Linear Model

Factor	Type	Levels	Values
Car	fixed	4	1 2 3 4
Driver	fixed	4	1 2 3 4
Gasoline	fixed	4	1 2 3 4
Road	fixed	4	1 2 3 4

Analysis of Variance for MPG, using Sequential SS for Tests

Source	DF	Seq SS	Adj SS	Seq MS	F	P
Car	3	16.500	16.500	5.500	2.75	0.214
Driver	3	6.500	6.500	2.167	1.08	0.475
Gasoline	3	111.500	111.500	37.167	18.58	0.019
Road	3	7.500	7.500	2.500	1.25	0.429
Error	3	6.000	6.000	2.000		
Total	15	148.000				

Minitab 输出中 Seq MS 列和表 16.31 中均方 MS 列一致. 计算的 F 值和表 16.31 的 F 值也相同. 汽车、司机、汽油及马路的 p -值分别为 0.214, 0.475, 0.019, 0.429. 由 p -值的性质可知, 在水平 0.01 及 0.05 下, 不同的汽车, 司机, 马路间不存在差异; 在水平 0.05 下, 汽油品牌间存在差异, 而在水平 0.01 下, 则不存在差异.

综合习题

16.18 证明 $\sum_j \alpha_j = 0$ (见本章式(15)).

证明 处理总体均值 μ_j 和总体均值 μ 之间有如下关系

$$\mu = \frac{1}{a} \sum_j \mu_j \quad (53)$$

因为 $\alpha_j = \mu_j - \mu$, 利用(53)式, 则

$$\sum_j \alpha_j = \sum_j (\mu_j - \mu) = \sum_j \mu_j - a\mu = 0 \quad (54)$$

16.19 证明(a)本章的(16)式, (b) 本章的(17)式.

证明 (a) 由定义知

$$V_W = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 = b \sum_{j=1}^a \left[\frac{1}{b} \sum_{k=1}^b (X_{jk} - \bar{X}_{j.})^2 \right] = b \sum_{j=1}^a S_j^2$$

其中 S_j^2 为第 j 个处理的样本方差, 样本的容量为 b . 所以

$$E(V_W) = b \sum_{j=1}^a E(S_j^2) = b \sum_{j=1}^a \left(\frac{b-1}{b} \sigma^2 \right) = a(b-1)\sigma^2$$

(b) 由定义知

$$V_B = b \sum_{j=1}^a (\bar{X}_{j.} - \bar{X})^2 = b \sum_{j=1}^a \bar{X}_{j.}^2 - 2b\bar{X} \sum_{j=1}^a \bar{X}_{j.} + ab\bar{X}^2 = b \sum_{j=1}^a \bar{X}_{j.}^2 - ab\bar{X}^2$$

因为 $\bar{X} = \frac{1}{a} \sum_j \bar{X}_{j.}$, 省略和式的下标 j , 则有

$$E(V_B) = b \sum E(\bar{X}_{j.}^2) - abE(\bar{X}^2) \quad (55)$$

对任一随机变量 U , $E(U^2) = \text{var}(U) + [E(U)]^2$, 其中 $\text{var}(U)$ 表示 U 的方差. 因此

$$E(\bar{X}_{j.}^2) = \text{var}(\bar{X}_{j.}) + [E(\bar{X}_{j.})]^2 \quad (56)$$

$$E(\bar{X}^2) = \text{var}(\bar{X}) + [E(\bar{X})]^2 \quad (57)$$

因为处理总体服从以 $\mu_j = \mu + \alpha_j$ 为均值的正态分布, 所以

$$\text{var}(\bar{X}_{j.}) = \frac{\sigma^2}{b} \quad (58)$$

$$\text{var}(\bar{X}) = \frac{\sigma^2}{ab} \quad (59)$$

$$E(\bar{X}_{j.}) = \mu_j = \mu + \alpha_j \quad (60)$$

$$E(\bar{X}) = \mu \quad (61)$$

利用结论(53)及(56)~(61)式, 可知

$$\begin{aligned} E(V_B) &= b \sum \left[\frac{\sigma^2}{b} + (\mu + \alpha_j)^2 \right] - ab \left(\frac{\sigma^2}{ab} + \mu^2 \right) \\ &= a\sigma^2 + b \sum (\mu + \alpha_j)^2 - \sigma^2 - ab\mu^2 \\ &= (a-1)\sigma^2 + ab\mu^2 + 2b\mu \sum \alpha_j + b \sum \alpha_j^2 + ab\mu^2 \\ &= (a-1)\sigma^2 + b \sum \alpha_j^2 \end{aligned}$$

16.20 证明本章的定理 1.

证明 如习题 16.19 所示,

$$V_W = b \sum_{j=1}^a S_j^2 \quad \text{或} \quad \frac{V_W}{\sigma^2} = \sum_{j=1}^a \frac{bS_j^2}{\sigma^2}$$

其中 S_j^2 是从第 j 个处理总体所抽取的容量为 b 的样本方差. 从第十一章(8)式可知 $\frac{bS_j^2}{\sigma^2} \sim \chi^2(b-1)$; 又因 S_j^2 相互独立, 所以由第十二章 χ^2 的可加性可知 $\frac{V_W}{\sigma^2} \sim \chi^2(a(b-1))$.

补充习题

单因素方差分析

16.21 为了判断五种不同品种小麦 A, B, C, D, E 产量间是否有差别而设计了一个试验. 每个品种分别种植在 4 块土地中, 产量(蒲式耳/英亩)见表 16.32. 假设土地的肥沃度一样, 且小麦品种随机地播种在各块土地中, 在显著性水平(a) 0.05, (b) 0.01 下判断产量间是否存在差异?

表 16.32

A	20	12	15	19
B	17	14	12	15
C	23	16	18	14
D	15	17	20	12
E	21	14	17	18

- 16.22 某公司希望检测四种不同类型的轮胎 A, B, C, D. 轮胎的寿命(由行驶的里程数决定), 见表 16.33 (单位: 千英里), 其中每种轮胎应用在随机选择的 6 辆汽车上. 在显著性水平(a) 0.05, (b) 0.01 下判断不同类型轮胎的寿命间是否存在显著差异?

表 16.33

A	33	38	36	40	31	35
B	32	40	42	38	30	34
C	31	37	35	33	34	30
D	29	34	32	30	33	31

- 16.23 某教师想检验三种教学方法 I、II、III 的效果. 他从班上随机抽取了 15 个学生, 每个被随机分在一小组, 共有 3 组, 每组 5 人, 实行一种教学方法. 最后对 15 名学生进行统一测试, 成绩见表 16.34. 在显著性水平(a) 0.05, (b) 0.01 下判断三种教学方法间是否存在差异?

表 16.34

方法 I	75	62	71	58	73
方法 II	81	85	68	92	90
方法 III	73	79	60	75	81

观测值数目不等时所作的修正

- 16.24 表 16.35 列出某种汽车使用不同品牌的汽油时, 每加仑所行驶的里程数. 在显著性水平(a) 0.05, (b) 0.01 下判断不同品牌的汽油间是否存在差异?

表 16.35

品牌 A	12	15	14	11	15
品牌 B	14	12	15		
品牌 C	11	12	10	14	
品牌 D	15	18	16	17	14
品牌 E	10	12	14	12	

表 16.36

数学	72	80	83	75	
自然科学	81	74	77		
英语	88	82	90	87	80
经济学	74	71	77	70	

- 16.25 某学期中, 一个学生不同学科的各次成绩见表 16.36. 在显著性水平(a) 0.05, (b) 0.01 下判断该学生在各门学科上是否存在显著差异?

双因素方差分析

- 16.26 某厂生产的产品是由 3 个工人在 3 台不同机器上生产的. 该厂厂长希望了解(a) 工人之间有无差别, (b) 机器间有无差异. 表 16.37 是每个工人在不同机器上一天所生产的产品数. 在显著性水平 0.05 下对此进行分析.

表 16.37

	工人		
	1	2	3
机器 A	23	27	24
机器 B	34	30	28
机器 C	28	25	27

表 16.38

	玉米品种			
	I	II	III	IV
区组 A	12	15	10	14
区组 B	15	19	12	11
区组 C	14	18	15	12
区组 D	11	16	12	16
区组 E	16	17	11	14

- 16.27 在显著性水平 0.01 下,求解习题 16.26.
- 16.28 四种不同玉米种子种植在 5 个区组中.每个区组分为 4 小块,每小块随机指定种植一种玉米,产量见表 16.38(单位:蒲式耳/英亩).在显著性水平 0.05 下判断是否由于(a)土质的不同,(b)玉米品种的不同引起产量的显著差异?
- 16.29 在显著性水平 0.01 下,求解习题 16.28.
- 16.30 假设在习题 16.22 中,每种轮胎的第一个观测值用在第一种汽车上,第二个观测值用在第二种汽车上,依次下去.在显著性水平 0.05 下判断(a)轮胎间有无差异,(b)汽车品牌间有无差异?
- 16.31 在显著性水平 0.01 下,求解习题 16.30.
- 16.32 假设在习题 16.23 中,每种教学方法下的第一个数据是某一所学校学生的成绩,第二个数据是另一所学校学生的成绩,依次下去.在显著性水平 0.05 下判断(a)教学方法间有无差异,(b)学校间有无差异?
- 16.33 为了检验美国成年女学生的头发颜色和身高与学生成绩是否相关,做了些调查,结果见表 16.39.其中数据表示即将毕业学生中成绩前 10% 的人数.在显著性水平 0.05 下对此试验进行分析.

表 16.39

	红色	金黄色	深棕色
高	75	78	80
中	81	76	79
矮	73	75	77

表 16.40

A	16	18	20	23
B	15	17	16	19
C	21	19	18	21
D	18	22	21	23
E	17	18	24	20

- 16.34 在显著性水平 0.01 下,求解习题 16.33.

有重复的双因素试验

- 16.35 假设习题 16.21 的试验在美国南部施行,表 16.32 的列表示四种不同肥料作用下的亩产量,表 16.40 是美国西部施行同样试验的结果.在显著性水平 0.05 下判断是否由于(a)肥料的不同,(b)所处地域的不同引起产量的差异?
- 16.36 显著性水平 0.01 下,求解习题 16.35.
- 16.37 表 16.41 给出一周各天内 4 个工人在 2 台不同类型机器 I、II 上所生产的产品数.在显著性水平 0.05 下判断(a)工人间是否存在显著差异,(b)机器间是否存在显著差异?

表 16.41

	机器 I					机器 II				
	周一	周二	周三	周四	周五	周一	周二	周三	周四	周五
工人 A	15	18	17	20	12	14	16	18	17	15
工人 B	12	16	14	18	11	11	15	12	16	12
工人 C	14	17	18	16	13	12	14	16	14	11
工人 D	19	16	21	23	18	17	15	18	20	17

拉丁方

- 16.38** 为了检验四种不同施肥处理 (A, B, C, D) 及横向和纵向上土壤的变化对玉米产量的影响进行了试验, 结果见表 16.42 的拉丁方, 其中的数据表示每单位面积玉米的产量. 在显著性水平 0.01 下检验 (a) 肥料间没有差异, (b) 土壤间不存在差异.
- 16.39** 在显著性水平 0.05 下, 求解习题 16.38.
- 16.40** 见习题 16.33, 假设引进了一个新因素——各个学生的出生地 E, M, W , 见表 16.43. 在显著性水平 0.05 下判断是否由于 (a) 身高的不同, (b) 头发颜色的不同, (c) 出生地的不同引起女学生成绩的显著差异?

表 16.42

C 8	A 10	D 12	B 11
A 14	C 12	B 11	D 15
D 10	B 14	C 16	A 10
B 7	D 16	A 14	C 12

表 16.43

E 75	W 78	M 80
M 81	E 76	W 79
W 73	M 75	E 77

希腊-拉丁方

- 16.41** 为了生产一种优质鸡饲料, 在原配方中增加了两种不同数量 (数量分为四个等级) 的化学药品, 生成 16 种新饲料配方. 第一种化学药品的不同数量用 A, B, C, D 表示, 第二种用 $\alpha, \beta, \gamma, \delta$ 表示. 根据雏鸡的初始重量 (W_1, W_2, W_3, W_4) 及品种的不同 (S_1, S_2, S_3, S_4) 把它们分为 16 组. 每组喂一种配方的饲料. 雏鸡单位时间内增加的重量见表 16.44 给出的希腊-拉丁方. 在显著性水平 0.05 下, 用方差分析的方法进行分析.

表 16.44

	W_1	W_2	W_3	W_4
S_1	$C_\gamma 8$	$B_\beta 6$	$A_\alpha 5$	$D_\delta 6$
S_2	$A_\delta 4$	$D_\alpha 3$	$C_\beta 7$	$B_\gamma 3$
S_3	$D_\beta 5$	$A_\gamma 6$	$B_\delta 5$	$C_\alpha 6$
S_4	$B_\alpha 6$	$C_\delta 10$	$D_\gamma 10$	$A_\beta 8$

- 16.42** 4 个公司 (C_1, C_2, C_3, C_4) 均生产四种不同类型的电缆 (T_1, T_2, T_3, T_4), 4 个工人 (A, B, C, D) 用 4 台机器 ($\alpha, \beta, \gamma, \delta$) 测量电缆的强度, 测得的平均强度见表 16.45 的希腊-拉丁方. 在显著性水平 0.05 下, 用方差分析的方法进行分析.

表 16.45

	C_1	C_2	C_3	C_4
T_1	$A_\beta 164$	$B_\gamma 181$	$C_\alpha 193$	$D_\delta 160$
T_2	$C_\delta 171$	$D_\alpha 162$	$A_\gamma 183$	$B_\beta 145$
T_3	$D_\gamma 198$	$C_\beta 212$	$B_\delta 207$	$A_\alpha 188$
T_4	$B_\alpha 157$	$A_\delta 172$	$D_\beta 166$	$C_\gamma 136$

表 16.46

A	3	5	4	4
B	4	2	3	3
C	6	4	5	5

综合习题

- 16.43** 表 16.46 给出了经过三种化学药品 A, B, C 处理过的铁锈厚度. 在水平 (a) 0.05, (b) 0.01 下判断处理间是否存在显著差异?
- 16.44** 一试验用来测量高、中、矮三级成年男学生的智商, 结果见表 16.47. 在水平 (a) 0.05, (b) 0.01 下判断

是否由于身高的差异而智商有所差异？

表 16.47

高	110	105	118	112	90	
矮	95	103	115	107		
中	108	112	93	104	96	102

16.45 证明本章的结论(10),(11)和(12)式.

16.46 不同智商的熟练工人与非熟练工人的成绩结果见表 16.48. 在水平 0.05 下判断是否由于(a) 熟练程度, (b) 智商的不同而引起得分的不同？

表 16.48

	成 绩		
	高智商	中等智商	低智商
熟练工人	90	81	74
非熟练工人	85	78	70

16.47 在显著性水平 0.01 下, 求解习题 16.46.

16.48 表 16.49 给出了来自不同地区, 不同智商的一组大学生的测验成绩. 在水平 0.05 下, 分析此表, 并陈述结论.

表 16.49

	得 分		
	高智商	中等智商	低智商
东部	88	80	72
西部	84	78	75
南部	86	82	70
中北部	80	75	79

16.49 在显著性水平 0.01 下, 求解习题 16.48.

16.50 对于习题 16.37, 你是否能断定一周内各天生产的产品数间存在显著差异? 请解释.

16.51 在方差分析的计算中, 从输入数据中加减一个适当的常数, 并不影响结论. 若对每个输入数据均乘以或除以某个定常数, 结论又会如何?

16.52 若观测值数目不等, 推导(24), (25)和(26)式.

16.53 假设习题 16.43 中表 16.46 的数据取自于美国东北地区, 而表 16.50 取自于西北地区. 在显著性水平 0.05 下, 判断是否由于(a) 化学药品的不同, (b) 地区的不同而引起铁锈厚度有所差异?

表 16.50

A	5	4	6	3
B	3	4	2	3
C	5	7	4	6

表 16.51

A	17	14	18	12
B	20	10	20	15
C	18	15	16	17
D	12	11	14	11
E	15	12	19	14

16.54 见习题 16.21, 16.35, 假设在美国东北部也进行了试验, 结果见表 16.51. 在显著性水平 0.05 下判断是否由于(a) 肥料的不同, (b) 地区的不同而引起产量方面有所差异?

- 16.55 在显著性水平 0.01 下,求解习题 16.44.
- 16.56 在显著性水平 0.01 下,对表 16.52 给出的拉丁方进行方差分析,并陈述结论.
- 16.57 设计一个试验,形成如表 16.52 所示的拉丁方.

表 16.52

因子 2	因子 1		
	B 16	C 21	A 15
	A 18	B 23	C 14
	C 15	A 18	B 12

- 16.58 在显著性水平 0.01 下,对表 16.53 给出的希腊-拉丁方进行方差分析,并陈述结论.

表 16.53

因子 2	因子 1			
	A _γ 6	B _β 12	C _δ 4	D _α 18
	B _δ 3	A _α 8	D _γ 15	C _β 14
	D _β 15	C _γ 20	B _α 9	A _δ 5
	C _α 16	D _δ 6	A _β 17	B _γ 7

- 16.59 设计一个试验,形成如表 16.53 所示的希腊-拉丁方.
- 16.60 描述在有重复的三因素试验中,怎样运用方差分析的方法.
- 16.61 设计一个试验,分析习题 16.60 的求解过程.
- 16.62 证明本章的(a) (30)式,(b) (31)~(34)式.
- 16.63 事实上,你能找到一个(a) 2×2 拉丁方,(b) 3×3 希腊-拉丁方吗? 请解释.

第十七章 非参数检验

引言

前几章所考虑的许多假设检验,均需要样本所依赖的总体的分布的某些假定.例如,第十六章的单因素方差分析要求总体服从正态分布,且标准差相等.

实际情况中,上述的假定不一定完全合理,或者在应用中对这些假定有怀疑.例如总体分布具有高度偏性的情形.因此,统计学家设计了许多与总体的分布及相关参数无关的检验方法,因此被称之为**非参数检验**.

非参数检验可用来代替复杂检验,尤其在处理非数值数据方面作用很大.例如由顾客对某种谷类食品的爱好程度划分等级而得到的数据.

符号检验

表 17.1 表示两种不同类型的机器(I、II)在连续 12 天内所生产的不合格螺栓数,假设机器每天的生产量相等.我们希望检验假设 H_0 : 机器之间无差异,即表中所列的机器生产的不合格品数的不同仅仅是随机误差所引起,因此可认为两组样本来自于同一总体.

表 17.1

天	1	2	3	4	5	6	7	8	9	10	11	12
机器 I	47	56	54	49	36	48	51	38	61	49	56	52
机器 II	71	63	45	64	50	55	42	46	53	57	75	60

对于这种成对出现的样本,有一种简单的非参数检验方法:**符号检验**.这种方法首先计算两种机器每天所生产的不合格品的差值,然后记下其差值的符号.例如:第一天其差值为 $47 - 71$,是负值,因此取负号.同理得到如下符号序列.

$$- - + - - - + - + - - - \quad (1)$$

(其中有 3 正 9 负).假如 $+$, $-$ 号出现的可能性一样,我们会得到 6 个 $+$, 6 个 $-$.检验 H_0 即相当于在投掷一枚硬币 12 次出现 3 次正面($+$)、9 次反面($-$)的结论下检验此硬币是否均匀.这就涉及到第七章所提到的二项分布.习题 17.1 说明在显著性水平 0.05 下,利用二项分布的双边检验,我们不能拒绝 H_0 ,即在此显著性水平下,两种机器间不存在差异.

注 1 若某天两机器生产出相同的不合格品,序列(1)中将出现 0.此时,我们可忽略 0 所对应的成对样本值,仅用 11 对观测值来代替 12 对观测值.

注 2 通过对样本进行连续性修正,可用正态分布近似表示二项分布(见习题 17.2).

尽管符号检验对成对数据特别适用(如表 17.1 所示),它也可以用来解决单个样本的问题(见习题 17.3、17.4).

Mann-Whitney U 检验

表 17.2 表示由两种不同的合金(I、II)所制造的电缆的强度.此表中有两组样本,对应于合金 I 收集了 8 个数据,合金 II 收集了 10 个数据.我们想利用这些数据判断样本间是否存在差异,或等价于两组样本是否取自于同一总体.尽管此问题可通过第十一章的 t -检验得到解决,一种称为 **Mann-Whitney U 检验**,或简称为 **U 检验**的方法也同样适用.此方法由下列步骤构成:

表 17.2

合金 I				合金 II				
18.3	16.4	22.7	17.8	12.6	14.1	20.5	10.7	15.9
18.9	25.3	16.1	24.2	19.6	12.9	15.2	11.8	14.7

第一步:合并两组数据,并按递增的顺序排列,然后根据各个数据在序列中的位置依次给他们编号(数据的编号称为该数据的秩),此例中从 1 到 18. 当有两个或两个以上样本值相等(称存在**结点**或者**结**)时,它们的秩均被赋予同一结点秩的均值. 例如若将表 17.2 中的数据 18.9 改为 18.3, 则有两个 18.3, 它们的秩分别为 12 和 13, 但因为它们相等, 所以用 $\frac{1}{2}(12 + 13) = 12.5$ 来作为他们共同的秩.

第二步:计算每组样本的秩和(此例中为合金 I 与合金 II 各自的秩和), 用 R_1, R_2 表示. 用 N_1, N_2 表示各个样本的容量, 为了计算方便, 若样本容量不等, 令 N_1 为较小样本容量(此例中 N_1 取 8), 则必有 $N_1 \leq N_2$, R_1 和 R_2 分别与 N_1 和 N_2 相对应. 秩和 R_1 和 R_2 之间的显著性差异则意味着样本间存在显著性差异.

第三步:为了检验秩和之间是否存在差异, 从样本 1 出发得到如下统计量

$$U = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 \quad (2)$$

在两样本间无差异时, U 的抽样分布对称, 且均值和方差可由如下公式计算:

$$\mu_U = \frac{N_1 N_2}{2}, \quad \sigma_U^2 = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} \quad (3)$$

若 N_1, N_2 均大于等于 8, 可证明 U 的分布渐近于正态分布. 因此

$$z = \frac{U - \mu_U}{\sigma_U} \quad (4)$$

渐近于标准正态分布. 通过附录 II, 我们可判断样本间是否存在显著性差异. 习题 17.5 说明在 0.05 水平下, 电缆间存在显著性差异.

注 3 利用样本 2, 我们也可得到如下统计量:

$$U = N_1 N_2 + \frac{N_2(N_2 + 1)}{2} - R_2 \quad (5)$$

它的抽样分布与(2)一致, 均值和方差计算公式如(3). 统计量(5)与(2)是相关的, 因为若 U_1 与 U_2 分别表示统计量(2)与(5), 则

$$U_1 + U_2 = N_1 N_2 \quad (6)$$

$$R_1 + R_2 = \frac{N(N + 1)}{2} \quad (7)$$

其中 $N = N_1 + N_2$. 证明见结论(7).

注 4 (2)中的统计量 U 表示在对所有样本值按递增顺序排列后, 样本 1 的值小于样本 2 的值的次数总和. 这就为我们提供了另一种计算统计量 U 的计数方法.

Kruskal-Wallis H 检验

U 检验是决定两组样本是否来自于同一总体的非参数检验. 当样本不是两组, 而是 k 组时, 检验 k 组样本是否取自于同一总体, 就要采用 **Kruskal-Wallis H 检验**, 简称为 **H 检验**.

此检验可描述如下:设有 k 组容量分别为 N_1, N_2, \dots, N_k 的样本, 令 $N = N_1 + N_2 + \dots + N_k$, 再设 k 组数据合并并以递增顺序排列后各组样本的秩分别为 R_1, R_2, \dots, R_k , 定义统计量

$$H = \frac{12}{N(N + 1)} \sum_{j=1}^k \frac{R_j^2}{N_j} - 3(N + 1) \quad (8)$$

可证明 H 的分布非常接近于自由度为 $k-1$ 的 χ^2 -分布, 这里假设 N_1, N_2, \dots, N_k 至少 ≥ 5 .

H 检验为单因素方差分析提供了一种非参数方法, 也可进行推广.

有结点时 H 检验的修正

若样本观测值中有太多的结, 统计量(8)的 H 值比应有的值小, 为此要进行修正. 修正后的值记为 H_c , 可通过修正因子

$$1 - \frac{\sum (T^3 - T)}{N^3 - N} \quad (9)$$

得到, 即 $H_c = \left(1 - \frac{\sum (T^3 - T)}{N^3 - N}\right)^{-1} \cdot H$, 其中 T 是每个结所对应的观测值的个数, 求 \sum 则表示对所有结点的 $T^3 - T$ 求和. 若没有结, 则 $T=0$, (9)式变为 1, 此时无须进行修正. 实际上, 一般情况下修正可被忽略.

随机性的游程检验

尽管单词“随机”在本书中已用过多次(例如“随机抽样”、“随机投掷一枚硬币”), 但前几章并未给出任何对随机性的检验. 游程论则提供了一种对随机性的非参数检验.

要理解什么是游程, 先考虑由两个变量 a, b 所构成的序列, 如

$$a a | b b b | a | b b | a a a a | b b b | a a a a | \quad (10)$$

例如在投掷一枚硬币中, 用 a 表示“正面”, b 表示“反面”; 或在机器生产的螺栓中, a 表示“不合格品”, b 表示“合格品”.

游程可以这样定义, 即包含在两个不同符号间的同一符号集. 从左到右观察序列(10), 共有 7 个游程, 各个游程间用“|”隔开, 第一个游程包含两个 a , 第二个游程含有 3 个 b , 第三个游程含有 1 个 a , 等等.

很显然, 在序列的随机性和其游程的个数之间应有某种关系存在. 再看序列

$$a | b | a | b | a | b | a | b | a | b | a | b | \quad (11)$$

它从 a 到 b , 再从 b 到 a , 如此排列下去, 因此是一个周期模型, 这样的模型几乎不可能认为有随机性存在, 此时我们认为游程过多.

另一方面, 序列

$$a a a a a a | b b b b b | a a a a a | b b b b | \quad (12)$$

又好像是一个趋势模型, 其中 a 和 b 彼此成群出现, 此时游程又过少, 这样的序列也不能认为是随机的.

由上可知, 一个序列游程过多或过少都会被看成是非随机的, 否则可认为是随机的. 为了检验序列的随机性, 我们考虑由 N_1 个 a 和 N_2 个 b ($N_1 + N_2 = N$) 任意排列的长度为 N 的序列. 从这些序列集我们得到一个抽样分布: 统计量 V ——每个序列的游程个数的抽样分布. 其均值和方差如下:

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1, \quad \sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)} \quad (13)$$

运用公式(13), 可以在给定显著性水平下, 对序列的随机性进行检验. 在 N_1 和 N_2 均不小于 8 的条件下, 可以证明统计量 V 的分布非常接近于正态分布. 因此

$$z = \frac{V - \mu_V}{\sigma_V} \quad (14)$$

渐近于标准正态分布. 可利用附录 II 进行检验.

游程检验的进一步应用

下面是游程检验的其他方面的应用.

1. 数值数据高于中位数和低于中位数的随机性检验. 为了判断数值数据(例如收集在样本中的数据)是否随机, 首先按照数据收集的先后次序把数据排列, 然后计算出数据的中位数, 再根据每一个原始数据是高于还是低于中位数分别用 a (高于) 和 b (低于) 来替换原始数据, 若某值与中位数相等, 则剔除出样本. 样本是否随机就可根据 a, b 序列是否随机来判断(见习题 17.20).

2. 样本所在总体的差异. 假设容量为 m 和 n 的两组样本, 其元素分别为 a_1, a_2, \dots, a_m 及 b_1, b_2, \dots, b_n , 为了判断两组样本是否取自同一总体, 首先合并两组样本, 使其成为一递增序列, 序列长度为 $m+n$. 若序列的某些值相同, 可通过随机方法(例如运用随机数)来对它进行排序. 若最后的序列是随机的, 可知样本间不存在差异, 即来自于同一总体, 否则来自于不同总体. 此检验可作为 Mann-Whitney U 检验的备用选择(见习题 17.21).

Spearman 秩相关

非参数方法也能用来检测两个变量 X, Y 的相关性. 若变量的值不是很精确时, 可以将每组数据按照大小、重要性等方式排序, 从 1 到 N 确定它们的秩, 用数据的秩代替数据的值来分析变量间的相关性. 若 X, Y 均按此方式排序, 秩相关系数或 Spearman 秩相关公式定义如下:

$$r_s = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} \quad (15)$$

其中 D 表示相应 X, Y 的值所对应秩的差值, N 是成对数据 (X, Y) 的个数.

习题及解答

符号检验

17.1 见表 17.1, 在显著性水平 0.05 下检验零假设 H_0 : 机器 I 和机器 II 产量间没有差异; 备择假设 H_1 : 机器 I 和机器 II 产量间存在差异.

解 假设随机变量 X 表示同一条件下, 投掷一枚均匀硬币 12 次正面出现的次数. $X = 0, 1, \dots, 12$. 图 17-1 为 X 的二项分布图(近似正态曲线). 从第七章有关知识知, $X = i$ 的概率为

$$P(X = i) = \binom{12}{i} \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{12-i}$$

由此可知

$$\begin{aligned} P(X = 0) &= 0.00024, & P(X = 1) &= 0.00293, \\ P(X = 2) &= 0.01611, & P(X = 3) &= 0.05371. \end{aligned}$$

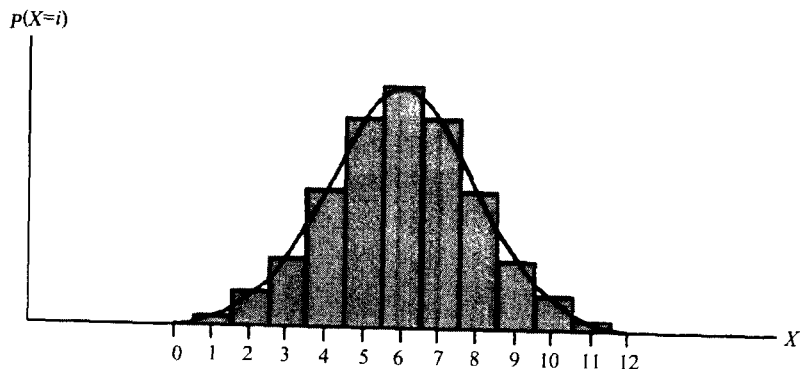


图 17-1

因为假设 H_1 表示机器间存在差异, 而非机器 I 产量高于机器 II, 所以采用双边检验. 给定显著性水平 0.05, 则每侧的尾概率为 $\frac{1}{2} \times 0.05 = 0.025$. 逐渐增加左侧尾概率直至其和超过 0.025, 即

$$P(X \leq 2) = 0.00024 + 0.00293 + 0.01611 = 0.01928$$

$$P(X \leq 3) = 0.00024 + 0.00293 + 0.01611 + 0.05371 = 0.07299$$

因为 $0.01928 < 0.025 < 0.07299$, 若正面出现的次数不超过 2 (或从对称角度来说, 正面出现次数不小于 10), 我们就拒绝零假设 H_0 , 但是此时正面出现的次数为 3 (本章序列 (1) 中 + 号的个数), 所以水平 0.05 下不能拒绝 H_0 , 即认为在此水平下, 机器产量间没有差异。

17.2 利用二项分布的正态逼近求解习题 17.1.

解 对于二项分布的正态逼近, 可构造统计量

$$z = \frac{X - \mu}{\sigma} = \frac{X - Np}{\sqrt{Npq}}$$

因为二项分布随机变量 X 是离散的, 而正态分布连续, 所以必须进行连续性修正 (例如正面次数 3 其实就是介于 2.5 到 3.5 间的一个数). 当 $X > Np$ 时, 用 $X - 0.5$ 替换 X ; 当 $X < Np$ 时, 相应增加 0.5. 此题中 $N = 12$, $\mu = Np = 12 \times 0.5$, $\sigma = \sqrt{Npq} = \sqrt{12 \times 0.5 \times 0.5} = 1.73$, 因此

$$z = \frac{(3 + 0.5) - 6}{1.73} = -1.45$$

因为 $-1.45 > -1.96$ (-1.96 为水平 0.05 下的双侧分位数), 可得到与习题 17.1 相同的结论。

注意, $P(z \leq -1.45) = 0.0735$, 和习题 17.1 中的 $P(X \leq 3) = 0.07299$ 非常接近。

- 17.3** PQR 公司宣称它生产的某种电池的寿命超过 250 小时. 一顾客希望验证此声明是否符合实际情况, 为此检测了该公司生产的 24 个电池, 结果见表 17.3. 假设样本是随机抽取的, 在显著性水平 0.05 下判断该公司的声明是否确切. 先用手算方法进行符号检验, 再用 Minitab 求解.

解 假设 H_0 表示该公司电池的寿命为 250 小时, H_1 表示寿命大于 250 小时. 要对 H_0, H_1 进行检验, 可用符号检验的方法. 先从表 17.3 的每个数据中减去 250, 记下各数的符号, 见表 17.4. 有 15 个加号, 9 个减号.

表 17.3

271	230	198	275	282	225	284	219
253	216	262	288	236	291	253	224
264	295	211	252	294	243	272	268

表 17.4

+	-	-	+	+	-	+	-
+	-	+	+	-	+	+	-
+	+	-	+	+	-	+	+

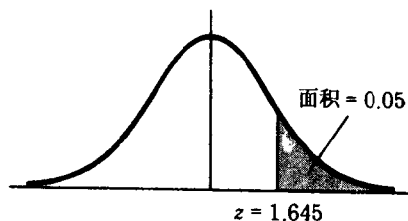


图 17-2

在水平 0.05 下, 运用单边检验, 若 z 值大于 1.645 (见图 17-2), 则拒绝 H_0 . 对 z 进行连续性修正, 则

$$z = \frac{(15 - 0.5) - 24 \times 0.5}{\sqrt{24 \times 0.5 \times 0.5}} = 1.02$$

因此在水平 0.05 下, 该公司的声明不能被证实。

用 Minitab 求解过程如下. 表 17.3 中的数据输入 Minitab 工作表中成为一列. 此列名为 Lifetime. 命令 STest 250 'Lifetime' 即可得到如下输出, 子命令 Alternative 1 表示进行右侧检验.

MTB> STest 250 'Lifetime';

SUBC> Alternative 1.

Sign Test for Median

Sign test of median = 250.0 versus > 250.0

	N	Below	Equal	Above	P	Median
Lifetime	24	9	0	15	0.1537	257.5

单边检验的 p -值为 0.1537. 它是零假设被拒绝的最小显著性水平, 因为 $0.05 < 0.1537$, 所以在水平 0.05 下不拒绝零假设。

17.4 表 17.5 给出了一次全国考试中 40 个学生的成绩. 在显著性水平 0.05 下, 检验假设 (a) 参加此次考试的学生成绩中位数为 66, (b) 参加此次考试的学生成绩中位数为 75. 先用手算方法进行符号检验, 再用 Minitab 求解.

表 17.5

71	67	55	64	82	66	74	58	79	61
78	46	84	93	72	54	78	86	48	52
67	95	70	43	70	73	57	64	60	83
73	40	78	70	64	86	76	62	95	66

解 (a) 从表 17.5 数据中均减去 66, 仅保留其符号, 得到表 17.6. 其中有 23 个加号, 15 个减号, 2 个 0. 删除 0 项, 则样本仅包含 38 个符号, 23 个加号, 15 个减号. 用正态分布的双边检验(左右尾概率各为 $\frac{1}{2} \times 0.05 = 0.025$), 则检验规则: 若 $-1.96 \leq z \leq 1.96$, 则接受假设, 否则拒绝假设.

表 17.6

+	+	-	-	+	0	+	-	+	-
+	-	+	+	+	-	+	-	-	-
+	+	+	-	+	+	-	-	-	+
+	-	+	+	-	+	+	-	+	0

因为

$$z = \frac{X - Np}{\sqrt{Npq}} = \frac{(23 - 0.5) - 38 \times 0.5}{\sqrt{38 \times 0.5 \times 0.5}} = 1.14$$

所以在水平 0.05 下, 接受中位数为 66 的假设.

我们也可用 15——减号的个数进行计算, 此时

$$z = \frac{(15 + 0.5) - 38 \times 0.5}{\sqrt{38 \times 0.5 \times 0.5}} = -1.14$$

结论相同.

(b) 从表 17.5 中减去 75 得到表 17.7, 其中有 13 个加号, 27 个减号. 因为

$$z = \frac{(13 + 0.5) - 40 \times 0.5}{\sqrt{40 \times 0.5 \times 0.5}} = -2.06$$

在水平 0.05 下, 拒绝中位数为 75 的假设.

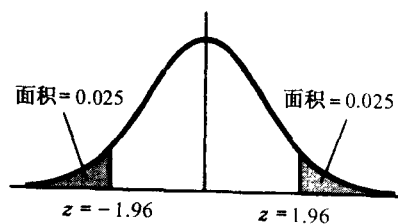


图 17-3

表 17.7

-	-	-	-	+	-	-	-	-	-
+	-	+	+	-	-	+	-	-	-
-	+	-	-	-	-	-	-	-	+
-	-	+	-	-	+	-	-	+	-

用此方法, 可得到中位数的 95% 置信区间(见习题 17.30).

用 Minitab 求解过程如下. 表 17.5 中的数据输入 Minitab 工作表中成为一列. 此列名为 Grade. 命令 STest 66 'Grade' 即可得到如下输出, 子命令 Alternative 0 表示进行双边检验.

MTB> STest 66 'Grade';

SUBC> Alternative 0.

Sign Test for Median

Sign test of median = 66.00 versus not = 66.00

	N	Below	Equal	Above	P	Median
Grade	40	15	2	23	0.2559	70.00

此检验的 p -值为 0.2559. 它是零假设被拒绝的最小显著性水平, 因为 $0.05 < 0.2559$, 所以在水平 0.05 下不拒绝零假设. 用 75 替换上述命令中的 66, 即可得到在假设中位数为 75 下的结论. 结果如下, 新的 p -值为 0.0385, 所以在水平 0.05 下, 拒绝中位数为 75 的零假设.

MTB> STest 75 'Grade';

SUBC> Alternative 0.

Sign Test for Median

Sign test of median = 75.00 versus not = 75.00

	N	Below	Equal	Above	P	Median
Grade	40	27	0	13	0.0385	70.00

Mann-Whitney U 检验

17.5 见表 17.2, 在水平 0.05 下, 判断由合金 I 与合金 II 制造的电缆间是否有差异? 先用手算方法进行 Mann-Whitney U 检验, 再用 Minitab 求解.

解 由本章前面的叙述得到第 1, 2, 3 步.

第一步: 合并 18 个样本值, 从小到大进行排列, 并在每个样本值下标上其秩. 具体见表 17.8.

表 17.8

10.7	11.8	12.6	12.9	14.1	14.7	15.2	15.9	16.1	16.4	17.8	18.3	18.9	19.6	20.5	22.7	24.2	25.3
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

第二步: 计算各个样本的秩和 R_1, R_2 , 得到表 17.9. 合金 I 的秩和 R_1 为 106, 合金 II 的秩和 R_2 为 65.

表 17.9

合金 I		合金 II	
电缆强度	秩	电缆强度	秩
18.3	12	12.6	3
16.4	10	14.1	5
22.7	16	20.5	15
17.8	11	10.7	1
18.9	13	15.9	8
25.3	18	19.6	14
16.1	9	12.9	4
24.2	17	15.2	7
	和 106	11.8	2
		14.7	6
			和 65

第三步: 因为合金 I 的样本容量较小, 所以 $N_1 = 8, N_2 = 10$. 相应的秩和为 $R_1 = 106, R_2 = 65$. 于是

$$U = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 = 8 \times 10 + \frac{8 \times 9}{2} - 106 = 10$$

$$\mu_U = \frac{N_1 N_2}{2} = \frac{8 \times 10}{2} = 40,$$

$$\sigma_U^2 = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} = \frac{8 \times 10 \times 19}{12} = 126.67$$

因此

$$\sigma_U = 11.25, \quad z = \frac{U - \mu_U}{\sigma_U} = \frac{10 - 40}{11.25} = -2.67$$

因为假设 H_0 是合金间不存在差异, 所以采用双边检验. 在水平 0.05 下, 判别规则如下: 若 $-1.96 \leq z \leq 1.96$, 则接受假设 H_0 , 否则拒绝假设 H_0 .

因为 $z = -2.67$, 拒绝假设 H_0 , 认为在水平 0.05 下, 合金 I、II 间存在差异.

Minitab 求解过程如下. 先把合金 I、II 的数据各输入到 Minitab 工作表的一列中, 命名为 AlloyI、AlloyII. 命令 Mann-Whitney 95.0 'AlloyI' 'AlloyII' 即可得到总体中位数之差的 95% 置信区间, Mann-Whitney 程序用来检验中位数相等的假设. Alternative 0 表示双边备择假设.

MTB> Mann-Whitney 95.0 'AlloyI' 'AlloyII';

SUBC> Alternative 0.

Mann-Whitney Confidence Interval and Test

AlloyI N=8 Median= 18.600

AlloyII N=10 Median= 14.400

Point estimate for ETA1-ETA2 is 4.800

95.4 Percent CI for ETA1-ETA2 is (2.000, 9.401)

W = 106.0

Test of ETA1=ETA2 vs ETA1 not = ETA2 is significant at 0.0088

Minitab 输出给出了每个样本的中位数, 总体中位数差的点估计以及置信区间, 第一个变量(此例中为 AlloyI)的秩和, 双边检验的 p -值=0.0088. 因为 $0.0088 < 0.05$ (显著性水平), 所以拒绝零假设, 即认为合金 I 制造的电缆强度大.

17.6 利用习题 17.5 的数据, 验证本章的结论(6)和(7).

解 (a) 由样本 1, 样本 2

$$U_1 = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 = 8 \times 10 + \frac{8 \times 9}{2} - 106 = 10$$

$$U_2 = N_1 N_2 + \frac{N_2(N_2 + 1)}{2} - R_2 = 8 \times 10 + \frac{10 \times 11}{2} - 65 = 70$$

故 $U_1 + U_2 = 10 + 70 = 80$, $N_1 N_2 = 8 \times 10 = 80$.

(b) 因为 $R_1 = 106$, $R_2 = 65$, 所以 $R_1 + R_2 = 106 + 65 = 171$,

$$\frac{N(N+1)}{2} = \frac{(N_1 + N_2)(N_1 + N_2 + 1)}{2} = \frac{18 \times 19}{2} = 171$$

17.7 利用合金 II 所对应的统计量 U 来求解习题 17.5.

解 对于合金 II 的样本,

$$U = N_1 N_2 + \frac{N_2(N_2 + 1)}{2} - R_2 = 8 \times 10 + \frac{10 \times 11}{2} - 65 = 70$$

因此

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{70 - 40}{11.25} = 2.67$$

此处的 z 值是习题 17.5 的 z 值的相反数(以正态分布的右尾代替左尾). 因为 z 值落在 $[-1.96, 1.96]$ 之外, 则结论与习题 17.5 相同.

17.8 某教授给两个班上心理学课. 上午班有 9 个学生, 下午班有 12 个学生. 所有学生同时进行期末考试, 成绩见表 17.10, 给定显著性水平 0.05, 你是否能认为上午班的成绩比下午班差? 先用手算方法进行 Mann-Whitney U 检验, 再用 Minitab 求解.

解

表 17.10

上午班	73	87	79	75	82	66	95	75	70			
下午班	86	81	84	88	90	85	84	92	83	91	53	84

第一步：表 17.11 给出了成绩与相应的秩，其中两个相同成绩 75 的秩为 $\frac{1}{2}(5+6)=5.5$ ，三个相同成绩 84 的秩为 $\frac{1}{3}(11+12+13)=12$ 。

表 17.11

53	66	70	73	75	75	79	81	82	83	84	84	84	85	86	87	88	90	91	92	95
1	2	3	4	5.5	5.5	7	8	9	10	12	12	12	14	15	16	17	18	19	20	21

第二步：由表 17.10 和表 17.11 得到表 17.12。知 $R_1=73$, $R_2=158$, $N=N_1+N_2=9+12=21$, 因此 $R_1+R_2=73+158=231$

$$\frac{N(N+1)}{2} = \frac{21 \times 22}{2} = 231 = R_1 + R_2$$

表 17.12

												秩和	
上午班	4	16	7	5.5	9	2	21	5.5	3			73	
下午班	15	8	12	17	18	14	12	20	10	19	1	12	158

第三步：

$$U = N_1 N_2 + \frac{N_1(N_1+1)}{2} - R_1 = 9 \times 12 + \frac{9 \times 10}{2} - 73 = 80$$

$$\mu_U = \frac{N_1 N_2}{2} = \frac{9 \times 12}{2} = 54$$

$$\sigma_U^2 = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} = \frac{9 \times 12 \times 22}{12} = 198$$

因此

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{80 - 54}{14.07} = 1.85$$

因我们希望在水平 0.05 下，检验备择假设 H_1 ：上午班学生成绩比下午班差对零假设 H_0 ：两班学生成绩无差异，所以考虑单边检验。由图 17-2 可知，判别规则如下：若 $z \leq 1.645$ ，则接受假设 H_0 ，否则拒绝假设 H_0 。

因为 $z = 1.85 > 1.645$ ，拒绝假设 H_0 ，认为在水平 0.05 下，上午班学生成绩比下午班差。在水平 0.01 下，不能得到此结论，因为此时临界值为 2.33，而 1.85 比 2.33 小。

Minitab 求解过程如下：先把上午班及下午班的数据各输入到 Minitab 工作表的一列中，命名为 Morning、Afternoon。命令 Mann-Whitney 95.0 'Morning' 'Afternoon' 即可得到总体中位数之差的 95% 置信区间，Mann-Whitney 程序用来检验中位数相等的假设。Alternative-1 表示左侧单边检验。
MTB>Mann-Whitney 95.0 'Morning' 'Afternoon';
SUBC>Alternative-1.

Mann-Whitney Confidence Interval and Test

Morning N = 9 Median = 75.00

Afternoon N = 12 Median = 84.50

Point estimate for ETA1-ETA2 is -9.00

95.7 Percent CI for ETA1 - ETA2 is (-15.00, 2.00)

W = 73.0

Test of ETA1 = ETA2 vs ETA1 < ETA2 is significant at 0.0350

The test is significant at 0.0348 (adjusted for ties)

Minitab 输出给出了每个样本的中位数, 总体中位数差的点估计以及置信区间, 第一个变量(此例中为上午班)的秩和, 单边检验的 p -值 $= 0.0348$. 因为 $0.0348 < 0.05$ (显著性水平), 所以拒绝零假设. 即认为上午班比下午班差. 当水平为 0.01 时, 不能得到此结论, 因为 p -值大于 0.01.

17.9 运用(a) 本章的公式(2), (b) 计数方法(详见本章注 4)计算表 17.13 的统计量 U .

解 (a) 把表 17.13 的数据按递增顺序排列, 并标出其秩, 见表 17.14. 用秩来替换表 17.13 相应数据, 得到表 17.15, 其中 $R_1 = 5, R_2 = 10$. 因为 $N_1 = 2, N_2 = 3$, 样本 1 的 U 值为

$$U = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 = 2 \times 3 + \frac{2 \times 3}{2} - 5 = 4$$

样本 2 的 U 值用同样方法得到为 $U = 2$.

表 17.13

样本 1	22 10
样本 2	17 25 14

表 17.14

数据	10 14 17 22 25
秩	1 2 3 4 5

表 17.15

		秩和
样本 1	4 1	5
样本 2	3 5 2	10

(b) 以 I, II 分别替换表 17.14 中的样本值, 来自样本 1 的用 I 替换, 其他用 II 替换. 则表 17.14 的第一行变为

数据	I	II	II	I	II
----	---	----	----	---	----

由此可看出

样本 2 第一个观测值前样本 1 的个数 = 1
 样本 2 第二个观测值前样本 1 的个数 = 1
 样本 2 第三个观测值前样本 1 的个数 = 2
 总数 = 4

因此对应于样本 1 的 U 值为 4.

同理可知

样本 1 第一个观测值前样本 2 的个数 = 0
 样本 1 第二个观测值前样本 2 的个数 = 2
 总数 = 2

因此对应于样本 2 的 U 值为 2.

因为 $N_1 = 2, N_2 = 3$, 因此这些值满足 $U_1 + U_2 = N_1 N_2$, 即 $4 + 2 = 2 \times 3 = 6$.

17.10 一总体包含三个元素: 7, 12, 15. 无放回地从该总体中抽取两个样本: 样本 1 包含一个元素; 样本 2 包含两个元素(两个样本正好用完总体).

(a) 写出 U 的抽样分布, 并绘制图形.

(b) 写出 U 的均值与方差.

(c) 利用本章公式(3)验证(b)的结论.

解 (a) 采用无放回抽样是为了避免结的出现. 此时有 $3 \times 2 = 6$ 种抽样方式, 如表 17.16 所示, 用秩 1, 2, 3 来替换 7, 12, 15. 表 17.16 的 U 值是对应于样本 1 的, 对于样本 2 来说, 分布也是一样.

表 17.16

样本 1	样本 2		U
7	12	15	2
7	15	12	2
12	7	15	1
12	15	7	1
15	7	12	0
15	12	7	0

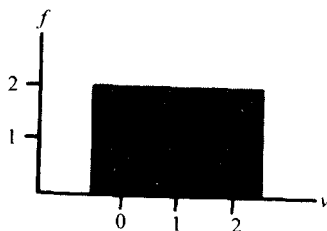


图 17-4

此分布的图形如图 17-4 所示, 其中 f 表示频数. U 的概率分布图也能如此绘制. 其中 $P(U=0) = P(U=1) = P(U=2) = \frac{1}{3}$. 其图形与图 17-4 相同, 只不过纵轴上的 1 与 2 分别用 $\frac{1}{6}$ 与 $\frac{1}{3}$ 替换.

(b) 从表 17.16 可得到均值与方差

$$\mu_U = \frac{2+2+1+1+0+0}{6} = 1$$

$$\sigma_U^2 = \frac{(2-1)^2 + (2-1)^2 + (1-1)^2 + (1-1)^2 + (0-1)^2 + (0-1)^2}{6} = \frac{2}{3}$$

(c) 由公式(3)

$$\mu_U = \frac{N_1 N_2}{2} = \frac{1 \times 2}{2} = 1$$

$$\sigma_U^2 = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} = \frac{1 \times 2 \times (1 + 2 + 1)}{12} = \frac{2}{3}$$

结果与(b)中结果一样.

- 17.11** (a) 写出习题 17.9 中 U 的抽样分布, 并绘制图形;
 (b) 绘制 U 的概率分布图;
 (c) 直接从(a)结论写出 U 的均值与方差;
 (d) 利用本章公式(3)验证(c)的结论.

解 (a) 此例中有 $5 \times 4 \times 3 \times 2 = 120$ 种抽样方式, 习题 17.9 的方法计算工作量太大, 为了简化计算, 把重点放在容量较小的样本($N_1=2$)及可能的秩和 R_1 . 当样本 1 包含秩(1, 2)时, 其秩和最小, $R_1=1+2=3$. 同理, 当样本 1 包含秩(4, 5)时, 秩和最大为 $R_1=4+5=9$. 因此可知 R_1 可从 3 变动到 9.

表 17.17 的第一列表示秩和 R_1 (从 3 到 9), 第二列表示相应的样本 1 的秩, 其和为 R_1 , 第三列给出了样本 1 的频数. 例如秩 $R_1=5$ 时, $f=2$. 因为 $N_1=2, N_2=3$, 所以

$$U = N_1 N_2 + \frac{N_1(N_1+1)}{2} - R_1 = 2 \times 3 + \frac{2 \times 3}{2} - R_1 = 9 - R_1$$

由此我们可知表 17.17 的第四列为相应的 U 值. 当 R_1 从 3 变动到 9 时, U 从 6 变动到 0. 抽样分布由第三列及第四列得到, 图 17-5 为其图.

(b) $U=9-R_1$ 的概率(即 $P(U=9-R_1)$)见表 17.17 的第五列, 通过频率得到. 频率为频数除以总频数(或 10)得到, $P(U=4) = \frac{2}{10} = 0.2$. 图 17-6 为概率分布图.

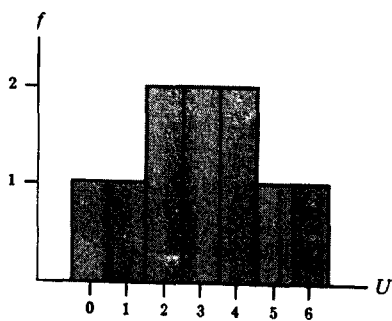


图 17-5

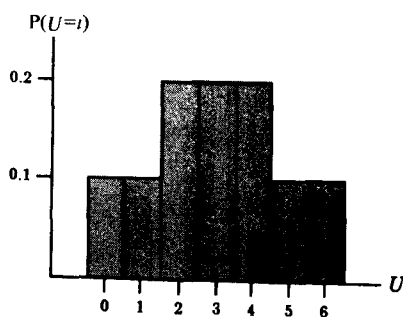


图 17-6

表 17.17

R_1	样本 1 的值	f	U	$P(U=9-R_1)$
3	(1, 2)	1	6	0.1
4	(1, 3)	1	5	0.1
5	(1, 4), (2, 3)	2	4	0.2
6	(1, 5), (2, 4)	2	3	0.2
7	(2, 5), (3, 4)	2	2	0.2
8	(3, 5)	1	1	0.1
9	(4, 5)	1	0	0.1

(c) 从表 17.17 的第三列与第四列知

$$\begin{aligned}\mu_U = \bar{U} &= \frac{\sum fU}{\sum f} = \frac{1 \times 6 + 1 \times 5 + 2 \times 4 + 2 \times 3 + 2 \times 2 + 1 \times 1 + 1 \times 0}{1 + 1 + 2 + 2 + 2 + 1 + 1} = 3 \\ \sigma_U^2 &= \frac{\sum f(U - \bar{U})^2}{\sum f} \\ &= \frac{1 \times (6-3)^2 + 1 \times (5-3)^2 + 2 \times (4-3)^2 + 2 \times (3-3)^2 + 2 \times (2-3)^2 + 1 \times (1-3)^2 + 1 \times (0-3)^2}{10} = 3\end{aligned}$$

另解

$$\begin{aligned}\sigma_U^2 &= \overline{U^2} - \bar{U}^2 \\ &= \frac{1 \times 6^2 + 1 \times 5^2 + 2 \times 4^2 + 2 \times 3^2 + 2 \times 2^2 + 1 \times 1^2 + 1 \times 0^2}{10} - 3^2 = 3\end{aligned}$$

(d) 由公式(3), 利用 $N_1=2, N_2=3$, 可知

$$\begin{aligned}\mu_U &= \frac{N_1 N_2}{2} = \frac{2 \times 3}{2} = 3, \\ \sigma_U^2 &= \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} = \frac{2 \times 3 \times 6}{12} = 3\end{aligned}$$

17.12 假定一序列从 1 变动到 N , 证明其秩和为 $\frac{N(N+1)}{2}$.

证明 令 R 为秩和, 则

$$R = 1 + 2 + 3 + \cdots + (N-1) + N \quad (16)$$

$$R = N + (N-1) + (N-2) + \cdots + 2 + 1 \quad (17)$$

其中(17)是通过改变(16)的求和顺序得到.(16), (17)式两边各项相加, 则有

$$R = \frac{N(N+1)}{2}$$

17.13 假设 R_1, R_2 是 U 检验中样本 1 和样本 2 的秩和, 证明 $R_1 + R_2 = \frac{N(N+1)}{2}$.

证明 假定样本中没有结点, 那么 R_1 必定为数集 $1, 2, \cdots, N$ 中某些数之和, R_2 是余下的数之和. 因此 $R_1 + R_2$ 一定为 $1, 2, \cdots, N$ 之和, 即 $R_1 + R_2 = 1 + 2 + \cdots + N = \frac{N(N+1)}{2}$.

Kruskal-Wallis H 检验

17.14 某公司想购买机器 A, B, C, D, E 之一. 令 5 个熟练工人分别在各个机器上工作相等时间以判断 5 台机器在产出上是否存在差异. 表 17.18 是相应的产品数. 在显著性水平(a)0.05, (b) 0.01 下检验假设 H_0 : 机器间不存在差异. 先用手算进行 Kruskal Wallis U 检验, 然后用 Minitab 求解.

解 因为有 5 个样本(A, B, C, D, E), 所以 $k=5$. 又因为每个样本包含 5 个数据, 所以 $N_1 = N_2 = N_3 = N_4 = N_5 = 5, N = N_1 + N_2 + N_3 + N_4 + N_5 = 25$. 把 25 个数据按递增顺序排列, 确定其秩, 用表 17.19 替换表 17.18, 它的最后一列为秩和. 从表 17.19 中可看出 $R_1 = 70, R_2 = 48.5, R_3 = 93, R_4 = 40.5, R_5 = 73$. 因此

表 17.18

A	68	72	77	42	53
B	72	53	63	53	48
C	60	82	64	75	72
D	48	61	57	64	50
E	64	65	70	68	53

表 17.19

						秩和
A	17.5	21	24	1	6.5	70
B	21	6.5	12	6.5	2.5	48.5
C	10	25	14	23	21	93
D	2.5	11	9	14	4	40.5
E	14	16	19	17.5	6.5	73

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{N_j} - 3(N+1)$$

$$= \frac{12}{25 \times 26} \left(\frac{70^2}{5} + \frac{48.5^2}{5} + \frac{93^2}{5} + \frac{40.5^2}{5} + \frac{73^2}{5} \right) - 3 \times 26 = 6.44$$

在显著性水平 0.05 下, H 服从自由度为 $k-1=4$ 的 χ^2 分布. 从附录 IV 知 $\chi_{0.95}^2(4) = 9.49$. 因为 $6.44 < 9.49$, 不能拒绝零假设, 即在水平为 0.05 时, 各机器间无差异. 当然水平为 0.01 时, 更不能拒绝零假设.

注意, 我们曾用方差分析的方法计算过此问题(见习题 16.8), 结论一致.

Minitab 求解过程如下. 先把数据以堆栈的形式输入到工作表中, 结构如下:

Row	Machine	Units
1	1	68
2	1	72
3	1	77
4	1	42
5	1	53
6	2	72
7	2	53
8	2	63
9	2	53
10	2	48
11	3	60
12	3	82
13	3	64
14	3	75
15	3	72
16	4	48
17	4	61
18	4	57
19	4	64
20	4	50
21	5	64
22	5	65
23	5	70
24	5	68
25	5	53

Minitab 命令 Kruskal-Wallis 'Units' 'Machine' 可得到如下输出

MTB> Kruskal-Wallis 'Units' 'Machine'.

Kruskal-Wallis Test

Kruskal-Wallis Test on Units

Machine	N	Median	Ave Rank	Z
1	5	68.00	14.0	0.34
2	5	53.00	9.7	-1.12
3	5	72.00	18.6	1.90
4	5	57.00	8.1	-1.66
5	5	65.00	14.6	0.54
Overall	25		13.0	

H=6.44 DF=4 P=0.168

H=6.49 DF=4 P=0.165(adjusted for ties)

注意, 结论中给出了两个 p -值, 其中一个是因为结点而进行调整之后的 p -值. 显然, 无论显著性水平为 0.05 还是 0.01, 机器间都不会存在显著差异, 因为两个 p -值都远大于 0.05 及 0.01.

17.15 对结点进行修正后, 再求解习题 17.14.

解 表 17.20 列出了有结点的观测值结点的个数. 例如, 48 出现了 2 次, 故 $T = 2$; 53 出现了 4 次, 故 $T = 4$. 计算所有的 $T^3 - T$, 然后求和, 得到 $\sum (T^3 - T) = 6 + 60 + 24 + 6 + 24 = 120$, 如表 17.20 所示. 又因为 $N = 25$, 所以修正因子为

$$1 - \frac{\sum (T^3 - T)}{N^3 - N} = 1 - \frac{120}{25^3 - 25} = 0.9923$$

修正后的 H 值为

$$H_C = \frac{6.44}{0.9923} = 6.49$$

此修正并未改变习题 17.14 所得的结论.

表 17.20

观测值	48	53	64	68	72	
结点的个数(T)	2	4	3	2	3	
$T^3 - T$	6	60	24	6	24	$\sum (T^3 - T) = 120$

17.16 表 17.21 是过去一年内教师、律师以及医生(均为随机选择)租借的 CD 数, 用 Minitab 的 Kruskal-Wallis H 检验, 在水平 0.01 下, 对零假设: 三种职业租借 CD 数目的分布相同进行检验.

表 17.21

教师	律师	医生
18	2	14
4	16	30
5	21	11
9	24	1
20	5	7
26	2	5
7	50	14
17	10	7
43	7	16
20	49	14
24	35	27
7	1	19
34	45	15
30	6	22
45	9	20
2	24	10
45	36	
9	50	
	44	
	3	

解 把数据按堆栈形式输入到工作表中, 各列分别记为 Rentals 和 Profession. 命令 Kruskal-Wallis 产生如下输出. 其中 p -值为 0.638, 此数值相当大, 它告诉我们三种职业租借 CD 数目在分布上不存在差异.

MTB>Kruskal-Wallis ‘rentals’ ‘Profession’;

Kruskal-Wallis Test

Profession	N	Median	Ave Rank	Z
1	18	19.00	28.9	0.47
2	20	18.50	28.7	0.44
3	16	14.00	24.4	-0.95
Overall	54		27.5	
H = 0.90 DF = 2 P = 0.638				
H = 0.90 DF = 2 P = 0.638 (adjusted for ties)				

随机性的游程检验

17.17 投掷一枚硬币 30 次, 得到由正面(H)与反面(T)组成的序列

H T T H T H H H T H H T T H T H T H H T H T T H T H H T H T

(a) 计算游程数 V .

(b) 在显著性水平 0.05 下, 检验假设 H_0 : 此序列是随机的.

先用手算得到随机性的游程检验, 然后再用 Minitab 求解.

解 (a) 用竖线将各个游程分离, 得到

H | T T | H | T | H H H | T | H H | T T | H | T | H | T | H H
| T | H | T T | H | T | H H | T | H | T |

可看出, 游程数 $V = 22$.

(b) 序列中出现 $N_1 = 16$ 个正面, $N_2 = 14$ 个反面, 由(a)知游程数 $V = 22$. 因此利用本章的公式(13)即可得到

$$\mu_V = \frac{2 \times 16 \times 14}{16 + 14} + 1 = 15.93,$$

$$\sigma_V^2 = \frac{2 \times 16 \times 14 \times (2 \times 16 \times 14 - 16 - 14)}{(16 + 14)^2(16 + 14 - 1)} = 7.175$$

或 $\sigma_V = 2.679$. 与游程数 $V = 22$ 相对应的 z 值为

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{22 - 15.93}{2.679} = 2.27$$

这是一个双边检验问题. 在水平 0.05 下, 若 $-1.96 \leq z \leq 1.96$, 则接受假设 H_0 , 认为序列是随机的; 否则拒绝 H_0 (见图 17-7). 因为计算得到的 z 值为 $2.27 > 1.96$, 故认为在水平 0.05 下, 投掷不是随机的. 此检验表明游程过多, 是一个循环模型.

假如进行连续性修正, 则 z 值为

$$z = \frac{(22 - 0.5) - 15.93}{2.679} = 2.08$$

结论与上述相同.

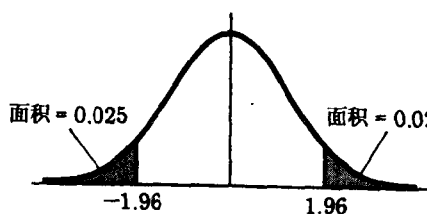


图 17-7

Minitab 求解过程如下. 数据按列输入, 正面用 1 来表示, 反面用 0 表示. 列用 Coin 来表示. Minitab 命令 Runs 'Coin' 即可得到如下输出

MTB>Runs 'Coin';

Runs Test

Coin

K = 0.5333

The observed number of runs = 22

The expected number of runs = 15.9333

16 Observations above K 14 below

The test is significant at 0.0235

K 值是数据列中 0 和 1 的均值. 高于或低于 K 的观测值数目分别为正面和反面出现的次数.

p -值为 0.0235. 它是零假设被拒绝的最小显著性水平, 因此在水平 0.05 下, 可看出零假设被拒绝.

17.18 某机器生产的 48 个工具按如下顺序排列. 其中 G 表示合格品, D 表示次品.

G G G G G G D D G G G G G G G
G G D D D D G G G G G G D G G G
G G G G G G D D G G G G G D G G

在显著性水平 0.05 下对序列进行随机性检验.

解 17.18 D 和 G 的个数分别为 $N_1=10$, $N_2=38$, 游程数 $V=11$. 均值和方差为

$$\mu_V = \frac{2 \times 10 \times 38}{10 + 38} + 1 = 16.83$$

$$\sigma_V^2 = \frac{2 \times 10 \times 38 \times (2 \times 10 \times 38 - 10 - 38)}{(10 + 38)^2 (10 + 38 - 1)} = 4.997$$

因此 $\sigma_V = 2.235$.

这是一个双边检验问题. 在水平 0.05 下, 若 $-1.96 \leq z \leq 1.96$, 则接受假设 H_0 , 认为序列是随机的; 否则拒绝 H_0 (见图 17-7). 因为与游程数 $V=11$ 相对应的 z 值为

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{11 - 16.83}{2.235} = -2.61$$

$-2.61 < -1.96$, 故在水平 0.05 下, 拒绝 H_0 .

此检验表明游程过少, 对于次品好像存在趋势模型. 对此序列进行的深一步的检验保证了此结论的成立.

- 17.19 (a) 写出由 3 个 a, 2 个 b 形成的所有可能序列, 并给出每个序列的游程数 V .
(b) 写出 V 的抽样分布, 并绘制图形.
(c) 写出 V 的概率分布, 并绘制图形.

解 17.19 (a) 3 个 a, 2 个 b 可以组成 $\binom{5}{2} = \frac{5!}{2!3!} = 10$ 个序列, 见表 17.22, 它还提供了相应的游程数.

(b) V 的抽样分布见表 17.23 (从表 17.22 得到), 其中 V 表示游程数, f 表示频数. 相应的图表见图 17-8.

(c) V 的概率分布见图 17-9, 由表 17.23 得到, 其中频率等于频数除以总频数 $2+3+4+1=10$ 的商. 例如 $P(V=5) = \frac{1}{10} = 0.1$.

表 17.22

序列	游程数 (V)
a a a b b	2
a a b a b	4
a a b b a	3
a b a b a	5
a b b a a	3
a b a a b	4
b b a a a	2
b a b a a	4
b a a a b	3
b a a b a	4

表 17.23

V	f
2	2
3	3
4	4
5	1

17.20 直接写出习题 17.19 游程数 V 的 (a) 均值, (b) 方差.

解 17.20 (a) 从表 17.22 知

$$\mu_V = \frac{2+4+3+5+3+4+2+4+3+4}{10} = \frac{17}{5}$$

另解

$$\mu_V = \frac{\sum fV}{\sum f} = \frac{2 \times 2 + 3 \times 3 + 4 \times 4 + 1 \times 5}{2+3+4+1} = \frac{17}{5}$$

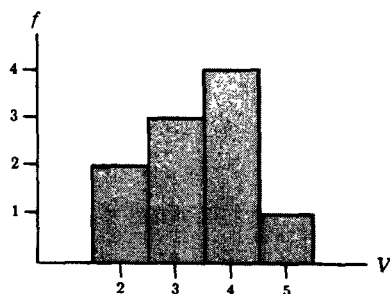


图 17-8

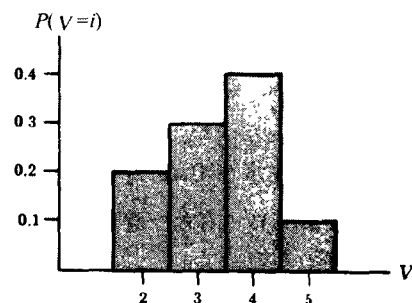


图 17-9

(b) 从表 17.22 知

$$\begin{aligned}\sigma_V^2 &= \frac{\sum f(V - \bar{V})^2}{\sum f} \\ &= \frac{1}{10} \left[2 \times \left(2 - \frac{17}{5} \right)^2 + 3 \times \left(3 - \frac{17}{5} \right)^2 + 4 \times \left(4 - \frac{17}{5} \right)^2 + 1 \times \left(5 - \frac{17}{5} \right)^2 \right] = \frac{21}{25}\end{aligned}$$

另解

$$\sigma_V^2 = \overline{V^2} - \bar{V}^2 = \frac{2 \times 2^2 + 3 \times 3^2 + 4 \times 4^2 + 1 \times 5^2}{10} - \left(\frac{17}{5} \right)^2 = \frac{21}{25}$$

17.21 利用本章公式(13)计算习题 17.20.

解 因为有 3 个 a, 2 个 b, 所以 $N_1 = 3, N_2 = 2$. 因此

(a)

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 = \frac{2 \times 3 \times 2}{3 + 2} + 1 = \frac{17}{5}$$

(b)

$$\begin{aligned}\sigma_V^2 &= \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)} \\ &= \frac{2 \times 3 \times 2 \times (2 \times 3 \times 2 - 3 - 2)}{(3 + 2)^2(3 + 2 - 1)} = \frac{21}{25}\end{aligned}$$

游程检验的进一步应用

17.22 见习题 17.3, 在水平 0.05 下, 判断 PQR 公司生产的电池寿命的样本序列是否随机? 若把表 17.3 的数据按行排列, 即第一个数据为 217, 第二个数据为 230, ..., 最后一个数据为 268. 先用手算方法进行随机性的游程检验, 再进行 Minitab 分析.

解 表 17.24 以递增次序对电池寿命进行了排列. 因为有 24 个输入数据, 所以中位数为中间两个数据(253 和 262)的均值, 即 $\frac{1}{2}(253 + 262) = 257.5$. 若表 17.3 的数据高于中位数, 则用 a 来表示, 若低于中位数, 用 b 表示, 这样就得到表 17.25, 其中有 12 个 a, 12 个 b. 因此 $N_1 = 12, N_2 = 12, N = 24, V = 15$, 故

$$\begin{aligned}\mu_V &= \frac{2N_1N_2}{N_1 + N_2} + 1 = \frac{2 \times 12 \times 12}{12 + 12} + 1 = 13 \\ \sigma_V^2 &= \frac{2 \times 12 \times 12 \times 264}{24^2 \times 23} = 5.739\end{aligned}$$

因此

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{15 - 13}{2.396} = 0.835$$

在水平 0.05 下进行双边检验, 若 $-1.96 \leq z \leq 1.96$, 则接受假设 H_0 , 认为序列是随机的; 否则拒绝 H_0 . 因为 0.835 未超出此范围, 因此我们认为样本是随机的.

表 17.24

198	211	216	219	224	225	230	236
243	252	253	253	262	264	268	271
272	275	282	284	288	291	294	295

表 17.25

a	b	b	a	a	b	a	b
b	b	a	a	b	a	b	b
a	a	b	b	a	b	a	a

Minitab 求解过程如下.数据按照收集的顺序按列输入,此列用 Lifetime 来表示.中位数由命令 Median c1 计算,等于 257.5. Minitab 命令 Runs 257.5 'Lifetime' 即可得到如下输出. p -值为 0.4038. 因此在水平 0.05 下,零假设不能被拒绝.

Lifetime

```
271    230    198    275    282    225    284    219    253
216    262    288    236    291    253    224    264    295
211    252    294    243    272    268
```

MTB>Median c1;

Column Median

Median of Lifetime = 257.50

MTB>Runs 257.5 'lifetime'.

Runs Test

Lifetime

K = 257.5000

The observed number of runs = 15

The expected number of runs = 13.0000

12 Observations above K 12 below

The test is significant at 0.4038

Cannot reject at alpha = 0.05

17.23 用随机性的游程检验求解习题 17.5.

解 从两个样本获得的数据已排在表 17.8 的第一行,用符号 a, b 分别表示取自样本 I、样本 II 的数据,则表 17.8 的第一行变为

b b b b b b b b a a a a a b b a a a

因为有 4 个游程,故 $N_1 = 8, N_2 = 10, V = 4$. 因此

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 = \frac{2 \times 8 \times 10}{18} + 1 = 9.889$$

$$\sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)} = \frac{2 \times 8 \times 10 \times 142}{18^2 \times 17} = 4.125$$

故

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{4 - 9.889}{2.031} = -2.90$$

若 H_0 表示合金之间无差异,它亦表示上述序列是随机的.在水平 0.05 下进行双边检验,若 $-1.96 \leq z \leq 1.96$,则接受假设 H_0 ,认为序列是随机的;否则拒绝 H_0 .因为 -2.90 超出此范围,因此我们拒绝 H_0 ,得到与习题 17.5 一致的结论.

若进行连续性修正,则

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{(4 + 0.5) - 9.889}{2.031} = -2.65$$

结果不变.

秩相关

17.24 表 17.26 列出了根据 10 个学生(按字母顺序排列)各自生物课的实验成绩与讲授课的成绩而构成的排序表(即秩).写出秩相关系数.

表 17.26

实验课	8	3	9	2	7	10	4	6	1	5
讲授课	9	5	10	1	8	7	3	4	2	6

解 每个学生实验课与讲授课的秩差 D 见表 17.27, 其中也给出了 D^2 和 $\sum D^2$. 因此

$$r_s = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} = 1 - \frac{6 \times 24}{10(10^2 - 1)} = 0.8545$$

表明在生物课的实验部分与讲授部分有明显的相关性.

表 17.27

秩差(D)	-1	-2	-1	1	-1	3	1	2	-1	-1
D^2	1	4	1	1	1	9	1	4	1	1
										$\sum D^2 = 24$

17.25 表 17.28 给出 12 对父亲与其大儿子(成年)的身高. 写出秩相关系数. 先用手算求解秩相关系数, 再用 Minitab 求解.

表 17.28

父亲身高(英寸)	65	63	67	64	68	62	70	66	68	67	69	71
儿子身高(英寸)	68	66	68	65	69	66	68	65	71	67	68	70

解 父亲身高按递增顺序排列为

$$62 \quad 63 \quad 64 \quad 65 \quad 66 \quad 67 \quad 67 \quad 68 \quad 68 \quad 69 \quad 70 \quad 71 \quad (18)$$

上述数组中第六、七位置上身高相同(皆为 67 英寸), 则他们的秩为 $\frac{1}{2}(6+7)=6.5$, 同理, 第八、九位置上身高相同(皆为 68 英寸), 则他们的秩为 $\frac{1}{2}(8+9)=8.5$. 因此父亲身高的秩排列为

$$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6.5 \quad 6.5 \quad 8.5 \quad 8.5 \quad 10 \quad 11 \quad 12 \quad (19)$$

同理, 儿子身高按递增顺序排列为

$$65 \quad 65 \quad 66 \quad 66 \quad 67 \quad 68 \quad 68 \quad 68 \quad 68 \quad 69 \quad 70 \quad 71 \quad (20)$$

上述数组中第六、七、八和九位置上身高相同(皆为 68 英寸), 则他们的秩为 $\frac{1}{4}(6+7+8+9)=7.5$. 同理, 因此儿子身高的秩排列为

$$1.5 \quad 1.5 \quad 3.5 \quad 3.5 \quad 5 \quad 7.5 \quad 7.5 \quad 7.5 \quad 7.5 \quad 10 \quad 11 \quad 12 \quad (21)$$

考虑(18)和(19), (20)和(21), 可用表 17.29 替换表 17.28. 表 17.30 给出了秩差 D , 以及 D^2 和 $\sum D^2$. 因此得知

$$r_s = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} = 1 - \frac{6 \times 72.50}{12(12^2 - 1)} = 0.7465$$

表 17.29

父亲身高的秩	4	2	6.5	3	8.5	1	11	5	8.5	6.5	10	12
儿子身高的秩	7.5	3.5	7.5	1.5	10	3.5	7.5	1.5	12	5	7.5	11

表 17.30

D	-3.5	-1.5	-1.0	1.5	-1.5	-2.5	3.5	3.5	-3.5	1.5	2.5	1.0
D^2	12.25	2.25	1.00	2.25	2.25	6.25	12.25	12.25	12.25	2.25	6.25	1.00
												$\sum D^2 = 72.50$

此结论和其他方法得到的相关系数一致(见习题 14.9, 14.14, 14.16 和 14.23).

Minitab 求解过程如下. 父亲身高和儿子身高在 Minitab 工作表中各为第一列与第二列, 分别记为 Father 和 Son.

Row	Father	Son
1	65	68
2	63	66
3	67	68
4	64	65
5	68	69
6	62	66
7	70	68
8	66	65
9	68	71
10	67	67
11	69	68
12	71	70

身高的秩分别输入到 c3 列及 c4 列.

MTB>rank c1 put into c3

MTB>rank c2 put into c4

MTB>print c1-c4

Row	Father	Son	C3	C4
1	65	68	4.0	7.5
2	63	66	2.0	3.5
3	67	68	6.5	7.5
4	64	65	3.0	1.5
5	68	69	8.5	10.0
6	62	66	1.0	3.5
7	70	68	11.0	7.5
8	66	65	5.0	1.5
9	68	71	8.5	12.0
10	67	67	6.5	5.0
11	69	68	10.0	7.5
12	71	70	12.0	11.0

用 Minitab 命令 correlation c3 c4 计算秩相关系数.

MTB>correlation c3 c4

Correlations(Pearson)

Correlation of c3 and c4 = 0.740, P-Value = 0.006

p -值 0.006 可用来对零假设(总体的秩相关系数为 0)及备择假设(总体的秩相关系数不为 0)进行检验. 因此我们认为父亲与儿子身高上存在相关性.

补充习题

符号检验

- 17.26 某公司宣称若在汽车油箱里加上他们生产的产品, 每加仑汽油行驶的里程数将会提高. 为了证明它的有效性, 挑选了 15 辆不同的汽车, 记录下其在增加该产品与不增加该产品下每加仑汽油行驶的里程数, 结论见表 17.31. 假定行驶条件相同, 在显著性水平(a) 0.05, (b) 0.01 下判断是否会由于增加该产品而引起行程的差异.

表 17.31

增加	34.7	28.3	19.6	25.1	15.7	24.5	28.7	23.5	27.7	32.1	29.6	22.4	25.7	28.1	24.3
不增加	31.4	27.2	20.4	24.6	14.9	22.3	26.8	24.1	26.2	31.4	28.8	23.1	24.0	27.3	22.9

- 17.27 在水平 0.05 下,对于习题 17.26,你是否能断定增加该产品后每加仑汽油行驶的里程数要比原来高?
- 17.28 某减肥俱乐部在广告上说,若按照它设计的特别课程去减肥,一个月体重将减少至少 6%.为了对此进行验证,36 个成人参加了此培训,其中有 25 个人达到广告上说的效果,6 人体重增加,其他人保持不变.在显著性水平 0.05 下判断此课程设计是否有效?
- 17.29 某公司培训部经理认为,若对其销售员工进行特别培训,公司的年销售额将会增加.为了对此进行验证,24 人参加了培训,其中 16 人的销售额提高,6 人下降,2 人保持不变.在显著性水平 0.05 下判断此特别培训是否增加了公司的销售额?
- 17.30 MW Soda 公司为了检验公众对两种品牌可乐 A 和 B 的相对喜欢程度,在全国 27 个地区进行了“品尝测验”.其中 8 个地区品牌 A 受欢迎,17 个地区品牌 B 受欢迎,其他地区两者皆可.在显著性水平 0.05 下,你能否断定品牌 B 比 A 更受欢迎?
- 17.31 随机从某公司生产的绳子中抽取了 25 根测量其断裂强度,结果见表 17.32.据此样本,在显著性水平 0.05 下分别检验该公司的声明,即绳子的断裂强度为(a) 25, (b) 30, (c) 35, (d) 40.

表 17.32

41	28	35	38	23
37	32	24	46	30
25	36	22	41	37
43	27	34	27	36
42	33	28	31	24

- 17.32 说明怎样计算习题 17.4 数据的 95% 置信限?
- 17.33 设计一试验,并用符号检验方法求解.

Mann-Whitney U 检验

- 17.34 XYZ 大学讲师 A 和 B 均担任某门基础化学的主讲教师,某次期末考试学生的成绩见表 17.33.在显著性水平 0.05 下检验假设:讲师 A 和 B 的学生的成绩不存在差异.

表 17.33

A	88	75	92	71	63	84	55	64	82	96				
B	72	65	84	53	76	80	51	60	57	85	94	87	73	61

- 17.35 见习题 17.8,在显著性水平 0.01 下,能否断定上午班学生的成绩比下午班差?
- 17.36 某农民想了解两个品种的小麦 I、II 产量上是否存在区别.表 17.34 展示了单位面积上各个品种小麦的产量.分别在显著性水平 (a) 0.05, (b) 0.01 下判断两个品种间是否存在差别?

表 17.34

小麦 I	15.9	15.3	16.4	14.9	15.3	16.0	14.6	15.3	14.5	16.6	16.0
小麦 II	16.4	16.8	17.1	16.9	18.0	15.6	18.1	17.2	15.4		

- 17.37 见习题 17.36,在显著性水平 0.05 下,此农民能否断定小麦 II 产量比小麦 I 高?
- 17.38 某公司希望了解两种品牌汽油 A 和 B 每加仑的行驶里程数是否有区别.表 17.35 给出两种品牌汽油每加仑的行驶里程数.在显著性水平 0.05 下,判断(a) 两个品牌间是否存在差异? (b) 品牌 B 是否

比品牌 A 要好?

表 17.35

A	30.4	28.7	29.2	32.5	31.7	29.5	30.8	31.1	30.7	31.8
B	33.5	29.8	30.1	31.4	33.8	30.9	31.3	29.6	32.8	33.0

- 17.39 能否用 U 检验来判断表 17.1 中的机器 I 和机器 II 之间是否存在差异? 请解释.
- 17.40 设计一试验, 并用 U 检验方法求解.
- 17.41 分别用 (a) 公式法, (b) 计数方法来计算表 17.36 数据的统计量 U .
- 17.42 以表 17.37 的数据替换表 17.36 的数据, 求解习题 17.41.

表 17.36

样本 1	15	25
样本 2	20	32

表 17.37

样本 1	40	27	30	56
样本 2	10	35		

- 17.43 某总体由 2, 5, 9 和 12 四个值组成. 从此总体中抽取了两个样本, 第一个样本包含一个值, 第二个样本由剩余的三个值组成.
- (a) 计算 U 的抽样分布并绘制图形.
- (b) 计算此分布的均值和方差, 分别用公式及直接计算.
- 17.44 证明 $U_1 + U_2 = N_1 N_2$.
- 17.45 在结点数分别为 (a) 1, (b) 2, (c) 任意数的情况下证明 $R_1 + R_2 = \frac{N(N+1)}{2}$.
- 17.46 若 $N_1 = 14, N_2 = 12, R_1 = 105$, 计算 (a) R_2 , (b) U_1 , (c) U_2 .
- 17.47 若 $N_1 = 10, N_2 = 16, U_2 = 60$, 计算 (a) R_1 , (b) R_2 , (c) U_1 .
- 17.48 在 N_1, N_2, R_1, R_2, U_1 和 U_2 中, 可由其余数值决定的数值中哪个最大? 给出证明.

Kruskal-Wallis H 检验

- 17.49 用某试验来检验 5 个不同品种的小麦 A, B, C, D 和 E 的产量是否有区别. 每一品种分别种植在 4 块地上. 产量 (蒲式耳/英亩) 见表 17.38. 假设土壤肥沃度相同且每个品种随机地种植在某块地中. 在显著性水平 (a) 0.05, (b) 0.01 下判断各品种小麦的产量间是否存在差别?

表 17.38

A	20	12	15	19
B	17	14	12	15
C	23	16	18	14
D	15	17	20	12
E	21	14	17	18

表 17.39

A	33	38	36	40	31	35
B	32	40	42	38	30	34
C	31	37	35	33	34	30
D	27	33	32	29	31	28

- 17.50 某公司想对四种轮胎 A, B, C 和 D 进行检验. 轮胎的寿命 (由行驶的里程决定, 单位: 千英里) 见表 17.39. 每种轮胎随机地应用在 6 辆性能相同的汽车上. 在显著性水平 (a) 0.05, (b) 0.01 下判断轮胎间是否存在显著差异?
- 17.51 某教师想对三种教学方法 I、II 和 III 进行检验. 为此, 他随机选择了三组学生 (每组 5 人) 进行试验, 每组学生采用一种教学方法. 最后进行统一测试, 成绩见表 17.40. 在显著性水平 (a) 0.05, (b) 0.01 下判断教学方法间是否存在差别?

表 17.40

方法 I	78	62	71	58	73
方法 II	76	85	77	90	87
方法 III	74	79	60	75	80

- 17.52 一个学生在某学期各门课成绩见表 17.41, 在显著性水平 (a) 0.05, (b) 0.01 下判断各门学科成绩间

是否存在差别?

表 17.41

数学	72	80	83	75
科学技术	81	74	77	
英语	88	82	90	87
经济学	74	71	77	70

17.53 用 H 检验, 求解(a) 习题 16.9, (b) 习题 16.21, (c) 习题 16.22.

17.54 用 H 检验, 求解(a) 习题 16.23, (b) 习题 16.24, (c) 习题 16.25.

随机性的游程检验

17.55 计算下列序列的游程数 V .

(a) A B A B B A A A B B A B

(b) H H T H H H T T T T H H T H H T H T

17.56 为了了解人们对某产品的喜恶(喜欢用 Y 表示, 不喜欢用 N 表示), 抽取了 25 个人进行试验, 结果见下表:

Y Y N N N N Y Y Y N Y N N Y N
N N N N Y Y Y Y N N

(a) 计算序列的游程数 V ;

(b) 在显著性水平 0.05 下判断回答序列是否随机?

17.57 对本章序列(10), (11)进行游程检验, 并写出结论.

17.58 (a) 写出由 2 个 a , 1 个 b 组成的所有序列, 并计算每一个序列的游程数 V .

(b) 计算 V 的抽样分布并绘制图形.

(c) 计算 V 的概率分布并绘制图形.

17.59 对习题 17.58, 分别用(a) 直接计算, (b) 公式的方法计算 V 的均值和方差.

17.60 在下列条件下求解习题 17.58 和 17.59.

(a) 2 个 a , 2 个 b ; (b) 1 个 a , 3 个 b ; (c) 1 个 a , 4 个 b .

17.61 在下列条件下求解习题 17.58 和 17.59.

(a) 2 个 a , 4 个 b ; (b) 3 个 a , 3 个 b .

游程检验的进一步应用

17.62 在显著性水平 0.05 下, 判断表 17.5 列出的 40 个成绩序列是否随机?

17.63 某股票连续 25 天的收盘价见表 17.42. 在显著性水平 0.05 下, 判断此价格是否随机?

表 17.42

10.375	11.125	10.875	10.625	11.500
11.625	11.250	11.375	10.750	11.000
10.875	10.750	11.500	11.250	12.125
11.875	11.375	11.875	11.125	11.750
11.375	12.125	11.750	11.500	12.250

17.64 $\sqrt{2}$ 若表示成小数形式, 它的前若干位为 1.41421 35623 73095 0488... 从数字的随机性考虑, 你能得到什么结论?

17.65 从数字的随机性考虑, 对于下列数值你能得到什么结论?

(a) $\sqrt{3} = 1.73205\ 08075\ 68877\ 2935\cdots$

(b) $\pi = 3.14159\ 26535\ 89793\ 2643\cdots$

- 17.66 用随机性的游程检验求解习题 17.30.
 17.67 用随机性的游程检验求解习题 17.32.
 17.68 用随机性的游程检验求解习题 17.34.

秩相关

- 17.69 某次比赛中,要求两名裁判根据其欣赏程度对 8 名选手排次序(从 1 到 8),结果见表 17.43.
 (a) 写出秩相关系数.
 (b) 判断两名裁判在他们判决上是否一致.

表 17.43

裁判一	5	2	8	1	4	6	3	7
裁判二	4	5	7	3	2	8	1	6

- 17.70 用秩相关方法求解:
 (a) 习题 14.26, (b) 习题 14.42, (c) 习题 14.46, (d) 习题 14.63.
 17.71 将第十四章的积-矩相关系数公式中的数据排秩后代替原数据,由此推导出秩相关系数.用两种方法求解同一问题来验证它.
 17.72 对分组数据能写出秩相关系数吗?请解释,并用实例来验证.

第十八章 时间序列分析

时间序列

一个**时间序列**是一组在特定时刻(一般是等时间间隔)上的观测值.例如,美国若干年内的钢铁的年产量,证券交易中某股票的日收盘价,城市气象局按小时发布的气温,某商店的月销售额等等.

从数学上看,时间序列根据变量 Y (温度、收盘价等)在不同时刻 t_1, t_2, \dots 的值 Y_1, Y_2, \dots 来表示,因此 Y 是关于 t 的函数,用符号表示即 $Y = F(t)$.

时间序列图

时间序列可通过构造变量 Y 对 t 的图表来表示.例如,图 18-1 就是反映 Desert Compact Discs And Cassettes Distribution 公司压缩光盘季销售额的时间序列图,此销售数据图涉及到 1994~1997 的各个季度.

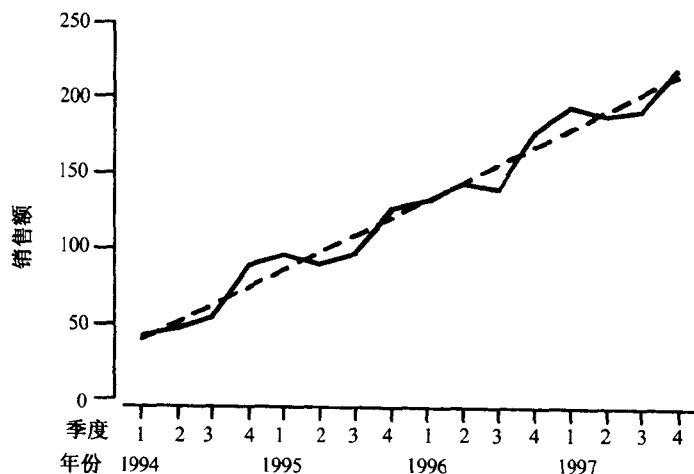


图 18-1 DCD CD 公司压缩光盘的季销售额

时间序列的特征运动

把时间序列图看作是一个描述随时间而移动的点图是很形象的,如图 18-1.许多情况下,它类似于一个粒子在力的作用下的路径图,它也可能是经济、社会学、精神或其他力作用下的质点的路径图.

时间序列的许多实例已证实了时间序列的某些特征运动,对这些运动的分析有很大价值.例如,对未来发展的预测等等.难怪许多工业及政府部门都很关注这门重要学科.

时间序列运动分类

时间序列的特征运动一般可分为四大类,常称之为时间序列的**运动成分或因素**.

1. 长期运动.此运动是指时间序列的长期发展趋势.在图 18-1 中,长期运动(或**长期变差、长期趋势**)由一条**趋势线**表示,即用虚线连接的线.对某些时间序列而言,**趋势曲线**可能更合适些.第十三章已通过最小二乘法拟合了趋势线及趋势曲线,本章后面还将介绍其他方法.

2. 循环运动或循环变差.此运动是指时间序列围绕趋势线或趋势曲线的长期振动或摆

动.这种循环运动,可能是**周期**的,也可能不是,也就是说,在等时间间隔中,时间序列有可能遵从完全相同的模型,也可能遵从不完全相同的模型.在商业和经济活动中,仅当同样性质的事件至少以一年间隔发生时,运动才能被认为是循环运动.循环运动的一个重要例子就是所谓的“经济循环”,它由经济繁荣、衰退、萧条、复苏构成.从图 18-1 的趋势线来看,循环运动表现得很明显.

3. 季节运动或季节变差.此运动是指在连续几年相应的月份或季度,时间序列模型相同或几乎完全相同.此运动是由于相同事件年年发生所致.例如圣诞节前百货商店的销售额剧增.季节运动在图 18-1 中可明显看到:4 年中每年第四季度的销售额最高.尽管在商业或经济理论中季节运动被看作是以**年度**为周期的,但其思想却可根据数据类型不同推广到任意时间间隔为周期的时间序列中(例如以天、小时、星期为周期).

4. 不规则运动或随机运动.此运动是指因为偶然事件而引起时间序列的随机变动.像洪水、罢工、选举都是偶然事件.尽管一般认为这样的事件只会引起短时间的变化,但它们也可能导致新的周期或其他运动.

时间序列分析

时间序列分析就是对目前各运动成分进行描述(一般是数学描述).为了对此有所了解,看图 18-2,它展示一个**理想**的时间序列.图 18-2(a)表示一条长期趋势线(而非趋势曲线),图 18-2(b)表示此长期趋势线上叠加一循环运动(可认为是周期的),图 18-2(c)表示图 18-2(b)上再叠加一个季节运动.若我们在图 18-2(c)上叠加不规则或随机运动,看起来和实际情况更相符.

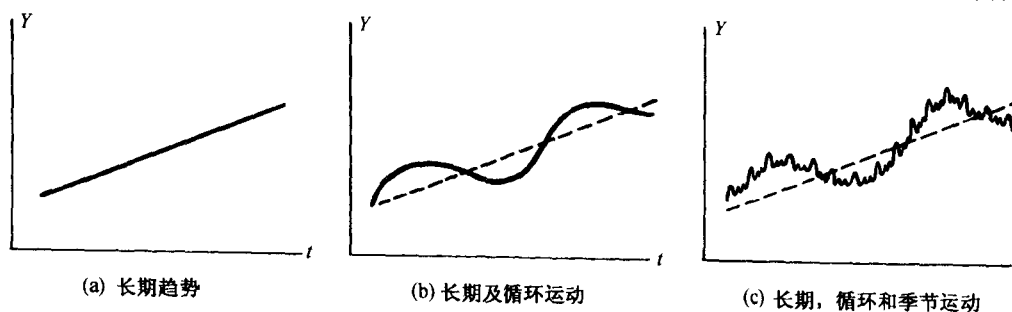


图 18-2

图 18-2 为我们提供了一种分析时间序列的方法.我们可假设时间序列变量 Y 是变量 T, C, S, I 的乘积(T, C, S, I 分别表示趋势、循环、季节、不规则运动),即

$$Y = T \times C \times S \times I = TCSI \quad (1)$$

时间序列分析相当于研究因素 T, C, S, I , 因此经常被认为是把一个时间序列**分解**成它的基本运动成分从而进行分析的过程.

有些统计学家更喜欢把 Y 看作是 $T + C + S + I$. 尽管我们在本章中用(1)所提供的分解进行分析,但和式分解也可得出相似的结论.实际上,至于用哪一种分解比较合适,还得看实际情况.

移动平均, 时间序列的平稳化

考虑数集

$$Y_1, Y_2, Y_3, \dots \quad (2)$$

上述序列相依的 N 项的算术平均可定义为 **N 阶移动平均**:

$$\frac{Y_1 + Y_2 + \dots + Y_N}{N}, \frac{Y_2 + Y_3 + \dots + Y_{N+1}}{N}, \frac{Y_3 + Y_4 + \dots + Y_{N+2}}{N}, \dots \quad (3)$$

序列(3)中的分子称为 **N 阶移动和**.

例 1 对于数字 2, 6, 1, 5, 3, 7 和 2 的 3 阶移动平均序列为

$$\frac{2+6+1}{3}, \frac{6+1+5}{3}, \frac{1+5+3}{3}, \frac{5+3+7}{3}, \frac{3+7+2}{3}, \text{或 } 3, 4, 3, 5, 4$$

通常, 把移动平均序列中的元素和原始数据按合适的对应关系排列起来, 如此例中为

原始数据 2, 6, 1, 5, 3, 7, 2

3 阶移动平均 3, 4, 3, 5, 4

移动平均序列中的每一元素为其正上方 3 个元素的均值.

按年或月收集的数据的 N 阶移动平均又称为 N -年移动平均, N -月移动平均, 例如 5-年移动平均, 12-月移动平均等, 也可用其他的时间单位.

移动平均有助于减少数据集的变差. 在时间序列中, 此性质经常用来减少不需要的波动, 因此这个过程称为**时间序列的平稳化**.

若在序列(3)中应用加权算术平均, 权重提前给出, 所得到的序列称为 N 阶**加权移动平均**.

例 2 若在例 1 中增加权重 1, 4, 1, 则 3 阶加权移动平均序列为

$$\frac{1 \times 2 + 4 \times 6 + 1 \times 1}{1+4+1}, \frac{1 \times 6 + 4 \times 1 + 1 \times 5}{1+4+1}, \frac{1 \times 1 + 4 \times 5 + 1 \times 3}{1+4+1}, \frac{1 \times 5 + 4 \times 3 + 1 \times 7}{1+4+1},$$
$$\frac{1 \times 3 + 4 \times 7 + 1 \times 2}{1+4+1}, \text{或 } 4.5, 2.5, 4.0, 4.0, 5.5.$$

趋势的估计

趋势可由下面几种方法来估计:

1. **最小二乘法**. 此方法见第十三章, 用此方法可找到一个与趋势线或趋势曲线相匹配的方程, 运用方程即可计算出趋势值 T .

2. **手画法**. 此方法仅通过观察时间序列图来拟合一条趋势线或趋势曲线, 从而估计 T . 但它主要依赖主观判断, 因此有明显的不利之处.

3. **移动平均法**. 通过运用适当阶的移动平均, 可消除循环, 季节, 不规则因素, 只留下趋势因素.

此方法的缺点: 序列开始和末尾的数据丢失, 在例 1 中开始有 7 个数据, 经过 3 阶移动平均之后仅剩 5 个; 移动平均可能造成原始数据没有的周期或其他运动; 移动平均还极易受极端值的影响. 为了克服这些缺点, 有时用带适当权重的加权移动平均来替代, 此时中间值的权重较大些, 而极端值的权重小些.

4. **半平均法**. 此方法把数据分成两部分(各部分的元素个数最好相等), 对各部分数据做平均, 则在时间序列图上得到两个点. 从这两个点即可得到趋势线, 从而直接计算趋势值 T .

尽管此方法运用起来简单, 如果不加区别应用, 结果很可能不尽如人意. 而且, 仅当趋势线是线性或近似于线性时才可应用. 它也可推广到把数据分为几部分, 每一部分的趋势线呈线性的情形.

季节变差的估计, 季节指数

为了判断(1)中的季节因素 S , 必须估计在一特定年度里时间序列的数据怎样从一个月变动到另一个月的变化情况. 表示一年各月变量的相对值的数据集就称为变量的**季节指数**. 例如, 若知道一月, 二月, 三月, ……的销售额是全年月平均销售额的 50%, 120%, 90%, …, 则 50%, 120%, 90%, …即可作为该年的季节指数. 一年季节指数的均值应为 100%, 即 12 个月的指数之和为 1200%.

计算季节指数的方法如下:

1. **百分数平均法**. 此方法中, 我们把每月的数据表示成年均值的百分数, 不同年但同一月的相应百分数进行平均(用均值或中位数皆可); 用均值最好避免可能的极端值. 最后得到的 12 个百分数即为季节指数. 若均值不是 100% (即和不是 1200%), 应进行调整, 也就是通过乘

以一适当因子得到。

2. 百分数趋势法.此方法中,把每月的数据表示成趋势值 T 的百分数,再用上述方法 1 得到相应的季节指数.若均值不是 100%,也用 1 中的方法进行调整。

注意从(1)中知 $Y/T = CSI$,对 Y/T 再进行平均即可得到季节指数.因为这些指数包含循环及不规则变差,因此此方法有一个很大的不利之处,特别当这些变差很大时。

3. 百分数移动平均法.此方法是先计算 12-月移动平均,因为其值落在相邻两月之间,所以再对其进行 2-月移动平均,这一结果称为 **12-月中心移动平均**。

然后再把每月的原始数据表示成相对应的 12-月中心移动平均的百分数,同一月份的百分数再进行平均,得到的就是季节指数.若其均值不是 100%,再进行相应调整。

由(1)可看出此法的逻辑推理. Y 的 12-月中心移动平均是为了消除季节因素 S 及不规则因素 I 的影响,等同于 TC 的值,因此,用原始数据 Y 除以 TC 即可得到 SI ,随后对相应月份进行平均是为了消除不规则因素 I ,从而得到季节指数 S 。

数据的消季节化

若用原始数据除以季节指数,所得数据称为**消季节化的或调整季节变差的数据**.这样的数据中仍包含有趋势,循环和不规则因素。

循环变差的估计

数据消去了季节因素后,也可进行趋势调整,只需用相应趋势值 T 去除数据即可.由(1)知,调整季节以及趋势因素的方法是用 Y 除以 ST ,从而得到 CI .用一适当的移动平均(例如 3, 5, 7 月,采用奇数是为了避免进行中心化)剔除不规则因素 I ,只留下周期因素 C .一旦循环变差被分离出来,就可进行详细研究.若时间序列遵循一个周期性或渐近周期性的循环,也可像构造季节指数一样构造**循环指数**。

不规则变差的估计

不规则变差可通过调整趋势、季节、循环变差得到估计.由(1)知,可用原始数据 Y 除以 T, S, C ,从而得到 I .实际上人们发现,不规则运动出现的机会较少,且一般服从正态分布。

数据的可比性

比较数据时必须始终小心谨慎,以便使比较具有公正性.例如,比较三月和二月数据时,必须注意到三月有 31 天,而二月有 28 或 29 天;比较不同年度二月份数据时,又要注意闰年二月有 29 天而非 28 天.另外,同一年或不同年不同月的工作日可能会因为节假日、罢工等因素而不同。

事实上,这些变差的调整并无一定的规律,调整依赖于调查员的谨慎。

预测

以上的方法可用于时间序列的预测.必须看到数学处理本身并不能解决所有问题,但若和研究者的感觉、经验、才智以及判断力结合起来,数学分析无论是在长期还是短期预测方面都有很大价值。

时间序列分析的基本步骤小结

1. 收集数据,确保数据的可靠性,牢记时间序列分析的最终目的.例如,假如想预测一给定的时间序列,获得相关的时间序列(还有其他信息)将有助于预测.必要时,还需对数据进行调整,如闰年、节假日等。

2. 绘制时间序列图,定性标注季节、长期趋势以及循环变差。

3. 构造长期趋势曲线(或直线),运用最小二乘法,手画法,移动平均法或半移动平均法求得适当的趋势值.
4. 若季节因素存在,计算季节指数,并对数据进行消季节化(即调整季节因素).
5. 调整趋势因素,使数据仅包含循环及不规则因素,进行 3, 5 或 7 阶移动平均消去不规则因素,只留下循环变差.
6. 绘制第 5 步中得到的循环变差图,注意可能存在周期或渐近周期.
7. 若想进行预测,则结合步骤 1~6 的结论,并运用其他可能得到的信息来进行.最后对各种误差来源和大小进行识别和估计.

习题及解答

时间序列的特征运动

18.1 对于如下的几种情况,你将想到时间序列的哪种特征运动?

- (a) 工厂发生火灾,从而使生产延误 3 周;(b) 经济的繁荣时期;(c) 某百货商店复活节后的销售额;(d) 由于人口的稳定增长而引起的小麦产量的增长;(e) 某市 5 年内的月降雨量.

解 特征运动分别为

- (a) 不规则运动,(b) 循环运动,(c) 季节运动,(d) 长期运动,(e) 季节运动.

移动平均,时间序列的平稳化

18.2 表 18.1 展示了 1985~1995 这 11 年每年美国的凶杀案总数(以千计).试构造(a) 5-年移动平均序列,(b) 4-年移动平均序列.

表 18.1

年份	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
凶杀案数	19.0	20.6	20.1	20.7	21.5	23.4	24.7	23.8	24.5	23.3	21.6

来源:美国联邦调查局.

解 (a) 见表 18.2, 第三列第一个移动和 101.9 是第二列第 1 到第 5 五个元素之和;第三列第二个移动和 106.3 是第二列第 2 到第 6 五个元素之和,依此类推.用 5 去除每一个移动和则得到第四列的移动平均.

表 18.2

年份	数据	5-年 移动和	5-年移动 平均
1985	19.0		
1986	20.6		
1987	20.1	101.9	20.38
1988	20.7	106.3	21.26
1989	21.5	110.4	22.08
1990	23.4	114.1	22.82
1991	24.7	117.9	23.58
1992	23.8	119.7	23.94
1993	24.5	117.9	23.58
1994	23.3		
1995	21.6		

表 18.3

年份	数据	4-年 移动和	4-年移动 平均
1985	19.0		
1986	20.6		
1987	20.1	80.4	20.100
1988	20.7	82.9	20.725
1989	21.5	85.7	21.425
1990	23.4	90.3	22.575
1991	24.7	93.4	23.350
1992	23.8	96.4	24.100
1993	24.5	96.3	24.075
1994	23.3	93.2	23.300
1995	21.6		

(b) 见表 18.3,作法与(a)一样,不同之处是(a)中对 5 个数据求和及平均,而(b)中对 4 个数据求和及平均.而且其移动和及移动平均位于两相邻年份的中间.若考虑偶数阶移动平均时,情形皆如此.

目前移动平均一般皆用统计软件来计算.表 18.2 及表 18.3 的结论可用 Minitab 直接求得.若利用下拉菜单 Stat→time series→moving average 即可得到任意阶的移动平均.4-阶移动平均计算结果如下:

MTB>print c1 c2

Data Display

Row	Murders	AVER1
1	19.0	*
2	20.6	*
3	20.1	*
4	20.7	20.100
5	21.5	20.725
6	23.4	21.425
7	24.7	22.575
8	23.8	23.350
9	24.5	24.100
10	23.3	24.075
11	21.6	23.300

注:AVER1 列所展示的数值与表 18.3 的第四列完全一致.

18.3 先用手算法构造习题 18.2 的一个 4-年中心移动平均,再用 Minitab 来求解.


解  **方法 I** 先计算一个 4-年移动平均,如习题 18.2(b);这些数值记录在相邻的两年之间,如表 18.4.若再对求出的移动平均计算一个 2-年移动和,结果就在所需的年份位置上,用 2 去除每一个结果,即可得到所需的 4-年中心移动平均(见第 5 列).

表 18.4

年份	数据	4-年移动平均	第三列的 2-年移动和	4-年中心移动平均 (第四列除以 2)
1985	19.0			
1986	20.6			
1987	20.1	20.100	40.825	20.413
1988	20.7	20.725	42.150	21.075
1989	21.5	21.425	44.000	22.000
1990	23.4	22.575	45.925	22.963
1991	24.7	23.350	47.450	23.725
1992	23.8	24.100	48.175	24.088
1993	24.5	24.075	47.375	23.688
1994	23.3	23.300		
1995	21.6			

方法 II 先计算一个 4-年移动和,如习题 18.2(b),数据列在两相邻年份之间,如表 18.5 所示.若再对所求得的数据计算 2-年移动和,则他们就会出现在所需的年份位置上,对第四列除以 $8(2 \times 4)$ 即可得到所需的移动平均.

表 18.5

年份	数据	4-年移动和	第三列的 2-年移动和	4-年中心移动平均 (第四列除以 8)
1985	19.0			
1986	20.6			
1987	20.1	80.4	163.3	20.413
1988	20.7	82.9	168.6	21.075
1989	21.5	85.7	176.0	22.000
1990	23.4	90.3	183.7	22.962
1991	24.7	93.4	189.8	23.725
1992	23.8	96.4	192.7	24.087
1993	24.5	96.3	189.5	23.688
1994	23.3	93.2		
1995	21.6			

Minitab 求解过程如下.把数据按列输入,打开下拉菜单 Stat→time series→moving average 即可得到中心移动平均,下面是输出结果.

MTB>print c1 c2

Data Display:

Row	Murders	AVER1
1	19.0	*
2	20.6	*
3	20.1	20.413
4	20.7	21.075
5	21.5	22.000
6	23.4	22.962
7	24.7	23.725
8	23.8	24.087
9	24.5	23.688
10	23.3	*
11	21.6	*

这与表 18.4、表 18.5 得到的结论一致.

- 18.4 对习题 18.3, 验证一个 4-年中心移动平均与一个 5-年加权移动平均一致, 其中权重分别为 1, 2, 2, 2, 1.

解 表 18.6 是 5-年加权移动平均表, 第三列的第一个数值为 $19.0 + 2 \times 20.6 + 2 \times 20.1 + 2 \times 20.7 + 21.5 = 163.3$, 第四列的第一个数值为 $163.3 \div 8 = 20.4125$ (四舍五入为 20.413), 第三列、第四列的其他值计算方法与此相同. 注: 表 18.6 的 5-年加权移动平均与表 18.4 的 4-年中心移动平均一致.

表 18.6

年份	数据	5-年加权移动和	5-年加权移动平均
1985	19.0		
1986	20.6		
1987	20.1	163.3	20.413
1988	20.7	168.6	21.075
1989	21.5	176.0	22.000
1990	23.4	183.7	22.963
1991	24.7	189.8	23.725
1992	23.8	192.7	24.088
1993	24.5	189.5	23.688
1994	23.3		
1995	21.6		

18.5 在图上绘制出习题 18.2(a)的移动平均序列及表 18.1 的原始数据序列.

解 原始数据在图 18-3 上以实线表示,移动平均用虚线表示.注意移动平均图比原始数据图光滑了许多,清晰地呈现了趋势线.但移动平均的一个缺点是时间序列的首尾数据丢失,当数据量不是很大时,后果可能比较严重.

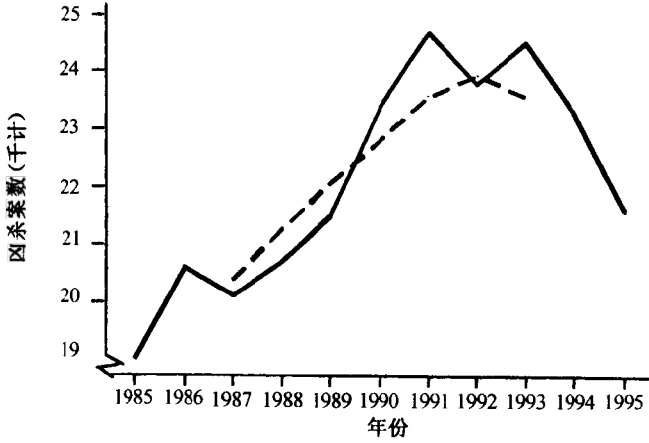


图 18-3

趋势估计

18.6 用半平均法求习题 18.2 的趋势值,分别采用(a) 均值, (b) 中位数.

解 (a) 把原始数据分成两部分(不考虑中间 1990 年的值),得到表 18.7. 然后计算每部分数据的均值.均值 20.38 对应到 1987 年,而 23.58 对应到 1993 年.

表 18.7

1985	19.0	1991	24.7
1986	20.6	1992	23.8
1987	20.1	1993	24.5
1988	20.7	1994	23.3
1989	21.5	1995	21.6
均值 = 20.38		均值 = 23.58	

过两点(1987, 20.38)、(1993, 23.58)的直线方程为 $y - 20.38 = 0.5333(x - 1987)$, 其中 x 表示年份, y 表示相应年份的凶杀案数. x 从 1985 年变动到 1995 年, 即可通过方程得到相应的 y 值, 从而得到趋势值, 见表 18.8.

表 18.8

年份	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
趋势值	19.31	19.85	20.38	20.91	21.45	21.98	22.51	23.05	23.58	24.11	24.65

(b) 1985 年到 1989 年数据的中位数为 20.6, 1991 年到 1995 年数据的中位数为 23.8. 过两点(1987, 20.6)、(1993, 23.8)的直线方程为 $y - 20.6 = 0.5333(x - 1987)$, 其中 x 表示年份, y 表示相应年份的凶杀案数. x 从 1985 年变动到 1995 年, 即可通过方程得到相应的 y 值, 从而得到趋势值, 见表 18.9.

表 18.9

年份	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
趋势值	19.53	20.07	20.60	21.13	21.67	22.20	22.73	23.27	23.80	24.33	24.87

若采用中位数进行计算,一般称此方法为**半中位数法**.

18.7 描述怎样利用(a)手画法,(b)移动平均法计算习题 18.2 的趋势值?

解 (a) 用手画法简单地构造一条直线或曲线,与图 18-3 很近似,然后从此线读取趋势值.

(b)从习题 18.5 可看出,5-年移动平均使时间序列平滑了许多.因此可用表 18.2 给出的移动平均值作为 1987 年到 1993 年的趋势值.

18.8 (a) 用 Minitab 拟合习题 18.2 的一条直线,用此最小二乘直线估计趋势值.

(b) 用 Minitab 拟合习题 18.2 的一条抛物线,用此最小二乘曲线估计趋势值.

解 (a) 拟合习题 18.2 的一条直线的 Minitab 输出如下:

Regression Analysis

The regression equation is

$$\text{Murder} = -817 + 0.422 \text{ Year}$$

Predictor	Coef	StDev	T	P
Constant	-817.3	263.8	-3.10	0.013
Year	0.4218	0.1326	3.18	0.011

$$S = 1.390 \quad R - Sq = 52.9\% \quad R - Sq(\text{adj}) = 47.7\%$$

把 1985 年到 1995 年的数据代入回归方程,就可得到趋势值,结果见表 18.10.

表 18.10

年份	凶杀案数	趋势值	残差
1985	19.0	20.00	-1.00
1986	20.6	20.42	0.18
1987	20.1	20.84	-0.74
1988	20.7	21.27	-0.57
1989	21.5	21.69	-0.19
1990	23.4	22.11	1.29
1991	24.7	22.53	2.17
1992	23.8	22.95	0.85
1993	24.5	23.37	1.13
1994	23.3	23.80	-0.50
1995	21.6	24.22	-2.62

(b) 拟合习题 18.2 的一条抛物线的 Minitab 输出如下:

Regression Analysis

The regression equation is

$$\text{Murder} = -411596 + 413 \text{ Year} - 0.104 \text{ YearSq}$$

Predictor	Coef	StDev	T	P
Constant	-411596	136581	-3.01	0.017
Year	413.3	137.3	3.01	0.017
YearSq	-0.10373	0.03449	-3.01	0.017

$$S = 1.010 \quad R - Sq = 77.9\% \quad R - Sq(\text{adj}) = 72.4\%$$

把 1985 年到 1995 年的数据代入回归方程,就可得到趋势值,结果见表 18.11.

表 18.11

年份	凶杀案数	趋势值	残差
1985	19.0	18.44	0.56
1986	20.6	19.80	0.80
1987	20.1	20.95	-0.85
1988	20.7	21.89	-1.19
1989	21.5	22.62	-1.12
1990	23.4	23.15	0.25
1991	24.7	23.46	1.24
1992	23.8	23.58	0.22
1993	24.5	23.48	1.02
1994	23.3	23.17	0.13
1995	21.6	22.66	-1.06

表 18.10 的残差平方和为 17.397, 表 18.11 的残差平方和为 8.165. 显然抛物线拟合效果更好, 趋势值更接近实际情况.

季节变差的估计; 季节指数

18.9 表 18.12 展示了从 1990 年 1 月到 1995 年 12 月美国的月新建住宅数量(以千计).

表 18.12

月份	1990	1991	1992	1993	1994	1995
1 月	99.2	52.5	71.6	70.5	76.2	84.5
2 月	86.9	59.1	78.8	74.6	83.5	81.6
3 月	108.5	73.8	111.6	95.5	134.3	103.8
4 月	119.0	99.7	107.6	117.8	137.6	116.9
5 月	121.1	97.7	115.2	120.9	148.8	130.5
6 月	117.8	103.4	117.8	128.5	136.4	123.4
7 月	111.2	103.5	106.2	115.3	127.8	129.1
8 月	102.8	94.7	109.9	121.8	139.8	135.8
9 月	93.1	86.6	106.0	118.5	130.1	122.4
10 月	94.2	101.8	111.8	123.2	130.6	126.2
11 月	81.4	75.6	84.5	102.3	113.4	107.2
12 月	57.4	65.6	78.6	98.7	98.5	92.8

来源: 美国人口调查局, 当前建筑报告.

- (a) 构造一个数据图.
(b) 用百分数平均法计算季节指数.

解 (a) 见图 18-4.

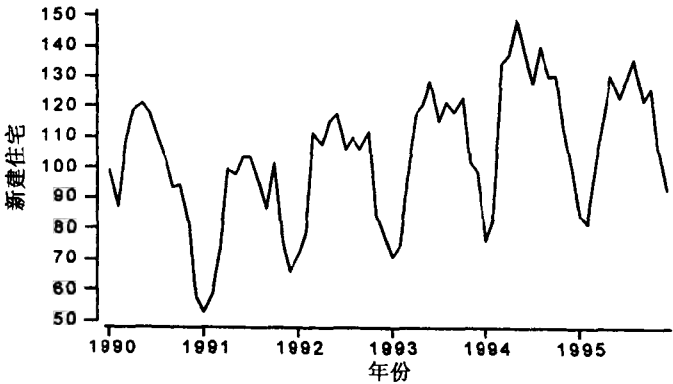


图 18-4 美国 1990~1995 年的新建住宅数

(b) 表 18.13 表示 1990~1995 年各年的新建住宅总数以及月均值。

用表 18.13 每年的月均值去除表 18.12 的相应年份的月数据, 结果以百分数形式列在表 18.14 中。例如: 1990 年 1 月对应的值为 $99.2/99.4 = 99.8\%$; 1991 年 1 月的数值由 $52.5/84.5 = 62.1\%$ 决定。表 18.14 的最后一列展示了每月的平均百分数。因为百分数之和为 1199.8, 所以需要进行调整, 即乘以因子 $1200/1199.8$, 使百分数之和达到 1200。但此调整对月百分数并没有显著影响。因此, 此列的数值就是我们所需要的季节指数。它表明, 从平均角度来看, 12 月、1 月和 2 月新建住宅最少; 而 5 月和 6 月最多。

表 18.13

年份	1990	1991	1992	1993	1994	1995
和	1192.6	1014.0	1199.6	1287.6	1457.0	1354.2
均值	99.4	84.5	100.0	107.3	121.4	112.9

表 18.14

月份	1990	1991	1992	1993	1994	1995	和	均值
1 月	99.8	62.1	71.6	65.7	62.8	74.8	436.8	72.8
2 月	87.4	69.9	78.8	69.5	68.8	72.3	446.7	74.5
3 月	109.2	87.3	111.6	89.0	110.6	91.9	599.7	99.9
4 月	119.7	118.0	107.6	109.8	113.3	103.5	672.0	112.0
5 月	121.8	115.6	115.2	112.7	122.6	115.6	703.5	117.2
6 月	118.5	122.4	117.8	119.8	112.4	109.3	700.1	116.7
7 月	111.9	122.5	106.2	107.5	105.3	114.3	667.6	111.3
8 月	103.4	112.1	109.9	113.5	115.2	120.3	674.3	112.4
9 月	93.7	102.5	106.0	110.4	107.2	108.4	628.2	104.7
10 月	94.8	120.5	111.8	114.8	107.6	111.8	661.2	110.2
11 月	81.9	89.5	84.5	95.3	93.4	95.0	539.6	89.9
12 月	57.7	77.6	78.6	92.0	81.1	82.2	469.3	78.2

18.10 用中位数代替均值, 计算习题 18.9 的季节指数。

解 表 18.14 中 1 月所在行的数据以递增顺序排列为 62.1, 62.8, 65.7, 71.6, 74.8 和 99.8, 其中位数为

$$(65.7 + 71.6)/2 = 68.7$$

其他月的中位数可用同样方法计算出来, 其结果见表 18.15。因为这些中位数之和为 1197.8, 所以可以通过因子 $1200/1197.8$ 进行调整, 调整后数据见表 18.15 的第三列, 它就是我们所需求的季节指数。实际上, 若用均值与中位数所得到的结果不同, 为了消除极值的影响, 最好采用中位数来计算。

表 18.15

月份	中位数	季节指数
1 月	68.7	68.8
2 月	71.1	71.2
3 月	100.5	100.7
4 月	111.6	111.8
5 月	115.6	115.8
6 月	118.2	118.4
7 月	109.7	109.9
8 月	112.8	113.0
9 月	106.6	106.8
10 月	111.8	112.0
11 月	91.4	91.6
12 月	79.9	80.0

18.11 用百分数趋势法计算习题 18.9 数据的季节指数.且用最小二乘法计算月趋势值.

解 从真实数据图(图 18-4)可看出,长期趋势可近似用一条直线来拟合,我们采用计算 1990 到 1995 年的月平均而非从表 18.12 的数据得到此直线,月平均见表 18.13,或表 18.16.

若把从 1990 年 1 月到 1995 年 12 月这 72 个月用 1 至 72 来编号,则 1990 年的月均值对应于时间点 6.5,1991 年的月均值对应于时间点 18.5,依次类推,具体结论见表 18.17.

表 18.16

年份	1990	1991	1992	1993	1994	1995
月均值	99.4	84.5	100.0	107.3	121.4	112.9

表 18.17

时间	6.5	18.5	30.5	42.5	54.5	66.5
月均值	99.4	84.5	100.0	107.3	121.4	112.9

用 Minitab 可得到表 18.17 数据的最小二乘直线.

Row	Y	X
1	99.4	6.5
2	84.5	18.5
3	100.0	30.5
4	107.3	42.5
5	121.4	54.5
6	112.9	66.5

MTB>regress 'Y' on 1 predictor 'X'

Regression Analysis

The regression equation is

$$Y = 88.1 + 0.442 X$$

当 X 从 1 变动到 72 时,应用上述回归方程,即可得到从 1990 年 1 月到 1995 年 12 月的所有趋势值,具体结果见表 18.18.

表 18.18

月份	1990	1991	1992	1993	1994	1995
1 月	88.5	93.6	98.6	103.7	108.8	113.8
2 月	88.9	94.0	99.1	104.1	109.2	114.3
3 月	89.4	94.4	99.5	104.6	109.6	114.7
4 月	89.8	94.9	99.9	105.0	110.0	115.1
5 月	90.2	95.3	100.3	105.4	110.5	115.5
6 月	90.6	95.7	100.8	105.8	110.9	116.0
7 月	91.1	96.1	101.2	106.2	111.3	116.4
8 月	91.5	96.5	101.6	106.7	111.7	116.8
9 月	91.9	97.0	102.0	107.1	112.2	117.2
10 月	92.3	97.4	102.4	107.5	112.6	117.6
11 月	92.7	97.8	102.9	107.9	113.0	118.1
12 月	93.2	98.2	103.3	108.4	113.4	118.5

用表 18.12 的数据除以表 18.18 中相应的趋势值,结果用百分数形式表示,见表 18.19.因为第八列之和为 1208.2,所以通过因子 1200/1208.2 进行调整即可得到第九列的季节指数;同理,第十列中位数之和为 1193,通过因子 1200/1193 调整后的数据见第十一列.

表 18.19

月份	1990	1991	1992	1993	1994	1995	均值	调整 均值	中位数	调整 中位数
1 月	112.1	56.1	72.6	68.0	70.0	74.3	75.5	75.0	71.3	71.7
2 月	97.8	62.9	79.5	71.7	76.5	71.4	76.6	76.1	74.1	74.5
3 月	121.4	78.2	112.2	91.3	122.5	90.5	102.7	102.0	101.8	102.3
4 月	132.5	105.1	107.7	112.2	125.1	101.6	114.0	113.2	110.0	110.6
5 月	134.3	102.5	114.9	114.7	134.7	113.0	119.0	118.2	114.8	115.5
6 月	130.0	108.0	116.9	121.5	123.0	106.4	117.6	116.8	119.2	119.9
7 月	122.1	107.7	104.9	108.6	114.8	110.9	111.5	110.7	109.8	110.4
8 月	112.3	98.1	108.2	114.2	125.2	116.3	112.4	111.6	113.3	113.9
9 月	101.3	89.3	103.9	110.6	116.0	104.4	104.3	103.5	104.2	104.8
10 月	102.1	104.5	109.2	114.6	116.0	107.3	108.9	108.2	108.3	108.9
11 月	87.8	77.3	82.1	94.8	100.4	90.8	88.9	88.3	89.3	89.8
12 月	61.6	66.8	76.1	91.1	86.9	78.3	76.8	76.3	77.2	77.7

18.12 用百分数移动平均法计算习题 18.9 的季节指数. 用 Minitab 辅助求解.

解 选择 Minitab 下拉菜单 Stat→Time series→Moving Average, 并选择长度为 12 的中心移动平均, 并保存移动平均. 分析结果见表 18.20.

表 18.20

月份	1990	1991	1992	1993	1994	1995
1 月	*	85.204	94.313	100.779	116.904	115.129
2 月	*	84.546	95.058	101.654	118.175	115.017
3 月	*	83.938	96.500	102.671	119.408	114.529
4 月	*	83.983	97.725	103.667	120.200	114.025
5 月	*	84.058	98.512	104.883	120.971	113.583
6 月	*	84.158	99.425	106.462	121.425	113.088
7 月	97.438	85.296	99.921	107.537	121.763	*
8 月	94.333	86.912	99.700	108.146	122.029	*
9 月	91.729	89.308	98.854	110.133	120.679	*
10 月	89.479	91.213	98.608	112.575	118.546	*
11 月	87.700	92.271	99.271	114.563	116.921	*
12 月	86.125	93.600	99.954	116.054	115.617	*

图 18-5 对应于表 18.20. 注意, 季节因素已经消除, 所以使图平滑了些. 现用 12-月中心移动平

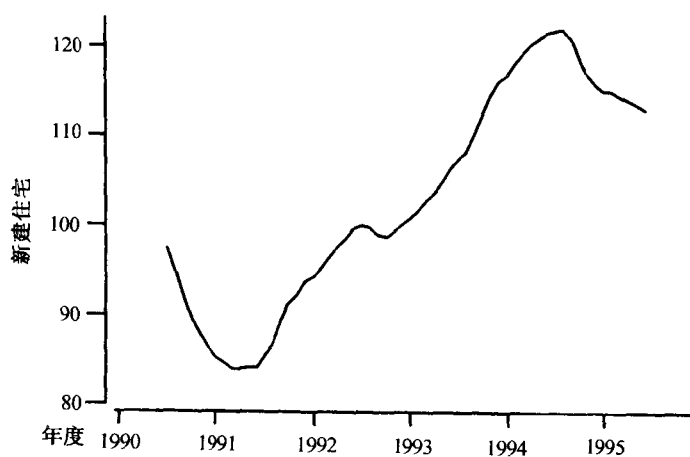


图 18-5 12-月中心移动平均

均去除相应的原始月数据,结果以百分数形式表示. Minitab 中,把 72 个新建住宅数据排在第一列,表 18.20 的 12-月中心移动平均排在第二列,选择命令 Let c3 = c1/c2. 第三列为表 18.21 中的值.

表 18.21

月份	1990	1991	1992	1993	1994	1995	均值	调整 均值	中位数	调整 中位数
1 月	*	61.6	75.9	70.0	65.2	73.4	69.2	69.7	70.0	70.8
2 月	*	69.9	82.9	73.4	70.7	70.9	73.6	74.1	70.9	71.8
3 月	*	87.9	115.6	93.0	112.5	90.6	99.9	100.7	93.0	94.2
4 月	*	118.7	110.1	113.6	114.5	102.5	111.9	112.7	113.6	115.1
5 月	*	116.2	116.9	115.3	123.0	114.9	117.3	118.2	116.2	117.7
6 月	*	122.9	118.5	120.7	112.3	109.1	116.7	117.6	118.5	120.0
7 月	114.1	121.3	106.3	107.2	105.0	*	110.8	111.6	107.2	108.6
8 月	109.0	109.0	110.2	112.6	114.6	*	111.1	111.9	110.2	111.6
9 月	101.5	97.0	107.2	107.6	107.8	*	104.2	105.0	107.2	108.6
10 月	105.3	111.6	113.4	109.4	110.2	*	110.0	110.8	110.2	111.6
11 月	92.8	81.9	85.1	89.3	97.0	*	89.2	89.9	89.3	90.4
12 月	66.6	70.1	78.6	85.0	85.2	*	77.1	77.7	78.6	79.6

季节指数见表 18.21, 因为第八列均值之和为 1191, 需要通过因子 1200/1191 进行调整, 结果见表 18.21 的第九列; 同理, 第十列中位数之和为 1185, 通过因子 1200/1185 修正结果见表 18.21 的第十一列.

18.13 用 Minitab 计算习题 18.9 数据的季节指数.

解 用 Minitab 下拉菜单 Stat→Time Series→Decomposition. 选择模型类型为 Multiplicative, 模型成分为 Trend plus seasonal, 选择 seasonals 并保存. Minitab 输出如下:

Time Series Decomposition

Data Starts

Length 72.0000

NMissing 0

Trend Line Equation

$Y_t = 87.7470 + 0.451756 * t$

Seasonal Indices

Period Index

1 0.708730

2 0.718379

3 0.943310

4 1.15295

5 1.17947

6 1.20045

7 1.08609

8 1.11415

9 1.08333

10 1.11307

11 0.903956

12 0.796113

若用 100 乘以指数值即可得到季节指数: 70.9, 71.8, 94.3, 115.3, 117.9, 120.0, 108.6, 111.4, 108.3, 111.3, 90.4 和 79.6.

Minitab 采用如下步骤计算季节指数.

第一步: 用最小二乘法拟合一条直线;

第二步: 用趋势因素去除相应数据, 对数据消除趋势;

第三步: 用长度等于季节周期的中心移动平均使消去趋势后的数据平滑, 对于月数据来说长度为 2;

第四步: 用移动平均数据去除消去趋势后的数据, 从而得到初步的季节指数;

第五步: 在每一个季节周期内, 计算初步季节指数的中位数. 调整中位数, 使它们的均值为 1.

下面就给出上述 5 步用 Minitab 软件的具体实现过程. 假定把表 18.12 的月新建住宅数据输入到工作表的一列, 而 1990 年 1 月到 1995 年 12 月 72 个月份以数字形式输入到另一列中, 例如, 1990 年 1 月用 1 来表示, 1990 年 2 月用 2 来表示, 依次类推, 直到 1995 年 12 月用 72 来表示. 最佳的最小二乘拟合直线的方程如下:

MTB> Regress 'Starts' 1 'time';

SUBC> Constant;

SUBC> Brief 1.

Regression Analysis

The regression equation is

Starts = 87.7 + 0.452 time

Predictor	Coef	StDev	T	P
Constant	87.747	4.741	18.51	0.000
time	0.4518	0.1129	4.00	0.000

S = 19.91 R-Sq = 18.6% R-Sq(adj) = 17.5%

趋势值通过回归方程 $\text{Starts} = 87.7 + 0.452 \text{ time}$ 得到, 其中 time 从 1 变动到 72, 具体结果见表 18.22.

表 18.22

. 月份	1990	1991	1992	1993	1994	1995
1 月	88.2	93.6	99.0	104.5	109.9	115.3
2 月	88.7	94.1	99.5	104.9	110.3	115.8
3 月	89.1	94.5	100.0	105.4	110.8	116.2
4 月	89.6	95.0	100.4	105.8	111.3	116.7
5 月	90.0	95.4	100.9	106.3	111.7	117.1
6 月	90.5	95.9	101.3	106.7	112.2	117.6
7 月	90.9	96.3	101.8	107.2	112.6	118.0
8 月	91.4	96.8	102.2	107.6	113.1	118.5
9 月	91.8	97.2	102.7	108.1	113.5	118.9
10 月	92.3	97.7	103.1	108.5	114.0	119.4
11 月	92.7	98.1	103.6	109.0	114.4	119.8
12 月	93.2	98.6	104.0	109.4	114.9	120.3

用表 18.22 的趋势值去除表 18.12 的数据值以消去趋势, 结论见表 18.23.

表 18.23

月份	1990	1991	1992	1993	1994	1995
1 月	1.125	0.561	0.723	0.675	0.693	0.733
2 月	0.980	0.628	0.792	0.711	0.757	0.705
3 月	1.218	0.781	1.117	0.906	1.212	0.893
4 月	1.329	1.050	1.072	1.113	1.237	1.002
5 月	1.345	1.024	1.142	1.138	1.332	1.114
6 月	1.302	1.078	1.163	1.204	1.216	1.050
7 月	1.223	1.074	1.044	1.076	1.135	1.094
8 月	1.125	0.978	1.075	1.132	1.237	1.146
9 月	1.014	0.891	1.033	1.096	1.146	1.029
10 月	1.021	1.042	1.084	1.135	1.146	1.057
11 月	0.878	0.770	0.816	0.939	0.991	0.895
12 月	0.616	0.665	0.756	0.902	0.858	0.771

表 18.23 的消去趋势的数据进行一个 12 阶中心移动平均,结果见表 18.24.

表 18.24

月份	1990	1991	1992	1993	1994	1995
1 月	*	0.910	0.951	0.964	1.063	0.999
2 月	*	0.898	0.954	0.968	1.070	0.994
3 月	*	0.887	0.964	0.973	1.076	0.985
4 月	*	0.883	0.972	0.978	1.079	0.976
5 月	*	0.879	0.975	0.985	1.081	0.969
6 月	*	0.877	0.981	0.996	1.082	0.961
7 月	1.075	0.885	0.983	1.003	1.082	*
8 月	1.036	0.899	0.977	1.006	1.081	*
9 月	1.003	0.920	0.965	1.020	1.066	*
10 月	0.974	0.935	0.958	1.038	1.043	*
11 月	0.949	0.940	0.960	1.051	1.024	*
12 月	0.926	0.949	0.961	1.060	1.008	*

用表 18.24 的移动平均数据去除表 18.23 的消去趋势的数据,得到初步的季节指数,结论见表 18.25.

表 18.25 的最后一列的季节指数,与下拉菜单 Stat→Time series→Decomposition 得到的结论一致.

表 18.25

月份	1990	1991	1992	1993	1994	1995	中位数	调整后的中位数
1 月	*	0.616	0.760	0.700	0.652	0.733	0.700	0.709
2 月	*	0.700	0.830	0.735	0.707	0.709	0.709	0.718
3 月	*	0.880	1.158	0.932	1.126	0.907	0.932	0.943
4 月	*	1.189	1.103	1.139	1.146	1.026	1.139	1.153
5 月	*	1.165	1.171	1.155	1.232	1.150	1.165	1.179
6 月	*	1.230	1.186	1.209	1.124	1.092	1.186	1.200
7 月	1.138	1.214	1.062	1.073	1.049	*	1.073	1.086
8 月	1.086	1.088	1.100	1.125	1.144	*	1.100	1.114
9 月	1.010	0.968	1.070	1.075	1.076	*	1.070	1.083
10 月	1.049	1.115	1.132	1.093	1.099	*	1.099	1.113
11 月	0.925	0.819	0.850	0.893	0.968	*	0.893	0.904
12 月	0.665	0.701	0.786	0.851	0.851	*	0.786	0.796

18.14 构造一个表格,比较习题 18.10,18.11,18.12 以及 18.13 所计算的季节指数.

解 见 表 18.26,它展示了由中位数计算得到的季节指数.

表 18.26

月份	百分数平均法	百分数趋势法	百分数移动平均法	Minitab 解
1 月	68.8	71.7	70.8	70.9
2 月	71.2	74.5	71.8	71.8
3 月	100.7	102.3	94.2	94.3
4 月	111.8	110.6	115.1	115.3
5 月	115.8	115.5	117.7	117.9
6 月	118.4	119.9	120.0	120.0
7 月	109.9	110.4	108.6	108.6
8 月	113.0	113.9	111.6	111.4
9 月	106.8	104.8	108.6	108.3
10 月	112.0	108.9	111.6	111.3
11 月	91.6	89.8	90.4	90.4
12 月	80.0	77.7	79.6	79.6

数据的消季节化

18.15 调整习题 18.9 数据的季节变差,即对数据进行消季节化,并展示如何用 Minitab 求得消季节化的数据.

解 要对数据进行季节变差的调整,就必须用习题 18.9 的每个原始数据除以相应的季节指数.例如,若采用习题 18.13 中 Minitab 计算出来的季节指数,就必须用 70.9%(即 0.709)去除 1 月份的所有数据.消去季节因素的数据见表 18.27.

表 18.27

月份	1990	1991	1992	1993	1994	1995
1 月	139.9	74.0	101.0	99.4	107.5	119.2
2 月	121.0	82.3	109.7	103.9	116.3	113.6
3 月	115.1	78.3	118.3	101.3	142.4	110.1
4 月	103.2	86.5	93.3	102.2	119.3	101.4
5 月	102.7	82.9	97.7	102.5	126.2	110.7
6 月	98.2	86.2	98.2	107.1	113.7	102.8
7 月	102.4	95.3	97.8	106.2	117.7	118.9
8 月	92.3	85.0	98.7	109.3	125.5	121.9
9 月	86.0	80.0	97.9	109.4	120.1	113.0
10 月	84.6	91.5	100.4	110.7	117.3	113.4
11 月	90.0	83.6	93.5	113.2	125.4	118.6
12 月	72.1	82.4	98.7	124.0	123.7	116.6

应用 Minitab 下拉菜单 Stat→Time series→Decomposition 即可得到表 18.27 所示的消去季节因素的数据.使用存储选择功能选择 seasonal adjusted data.表 18.27 的数据将表示成工作表的一列.

- 18.16 (a) 绘制习题 18.15 的消去季节因素的数据图;
(b) 把此图与习题 18.9(a)中的图 18-4 进行比较.

解 (a) 见图 18-6.

(b) 进行季节调整后的数据图展示了一个长期趋势, 从 1991 年开始相当接近于线性. 若把习题 18.9 的数据用 $Y = TCSI$ 表示, 则图 18-6 就是变量 $Y/S = TCI$ 关于时间的图形, 因此包含有长期趋势, 循环以及不规则运动.

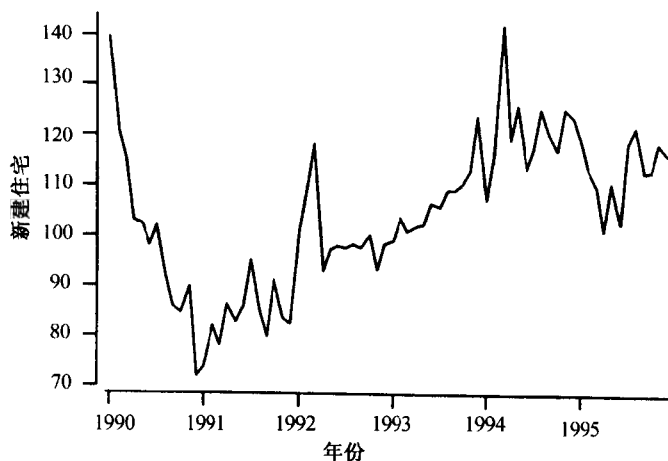


图 18-6 消去季节因素的数据

循环及不规则变差的估计

18.17 对习题 18.15 的数据进行趋势修正.

解 为了消除表 18.27 数据的趋势, 用表 18.27 的每一个数据除以相应的月趋势值 (任何一种方法计算出来的皆可). 此题中我们采用表 18.20 给出的月趋势值, 结论见表 18.28. 例如, 要得到 1990 年 7 月的数据, 用表 18.27 的相应值 102.4 除以 97.4 (见表 18.20) 即可, 结果为 $102.4/97.4 = 105.1\%$. 其他值的计算方法一样. 如同涉及到移动平均的任何方法一样, 此方法的缺点是时间序列首尾的数据丢失.

表 18.28

月份	1990	1991	1992	1993	1994	1995
1 月	*	86.9	107.1	98.6	92.0	103.5
2 月	*	97.3	115.4	102.2	98.4	98.8
3 月	*	93.3	122.6	98.7	119.3	96.1
4 月	*	103.0	95.5	98.6	99.3	88.9
5 月	*	98.6	99.2	97.7	104.3	97.5
6 月	*	102.4	98.8	100.6	93.6	90.9
7 月	105.1	111.7	97.9	98.8	96.7	*
8 月	97.8	97.8	99.0	101.1	102.8	*
9 月	93.8	89.6	99.0	99.3	99.5	*
10 月	94.5	100.3	101.8	98.3	98.9	*
11 月	102.6	90.6	94.2	98.8	107.3	*
12 月	83.7	88.0	98.7	106.8	107.0	*

18.18 (a) 绘制习题 18.17 的数据图.

(b) 解释此图的意义.

解 (a) 从习题 18.17 的数据中均减去 100%, 再对偏差数据进行绘制, 结果见图 18-7.

(b) 原始数据用 $Y = TCSI$ 表示. 对季节变差 (见习题 18.15) 的调整相当于在等式的两边同除以季节指数 S , 即 $Y/S = TCI$. 接下来对趋势的调整也就是对 $Y/S = TCI$ 的两边同除以 T , 即 $Y/ST =$

CI. 对上式两边同减去 100 个百分点, 可得 $(Y/ST) - 100 = CI - 100$. 因此图 18-7 的因变量为 $(Y/ST) - 100$, 自变量为时间 t . 图 18-7 从理论上说只包含循环及不规则运动 C 和 I . 图 18-7 的大部分变差低于 5%. 相距一年半左右的大变差暗示着可能存在周期因素, 但要确证它, 还需要进行大量地观测.

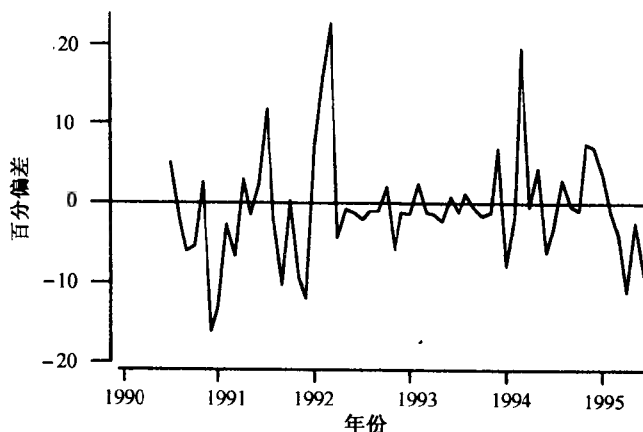


图 18-7 循环及不规则变差

- 18.19 (a) 求习题 18.17 数据的 3-月及 7-月移动平均.
 (b) 绘制(a)中移动平均的数据图.
 (c) 解释此图.

解 (a) 用 Minitab 计算出 3-月及 7-月移动平均, 减去 100 个百分点后的数据见表 18.29. 3-月移动平均位于表的上半部分, 而 7-月移动平均位于下侧.

表 18.29

月份	1990	1991	1992	1993	1994	1995
1 月	*	-10.7	3.5	-0.2	-0.9	3.1
2 月	*	-7.5	15.0	-0.2	3.2	-0.5
3 月	*	-2.1	11.2	-0.2	5.7	-5.4
4 月	*	-1.7	5.8	-1.7	7.6	-5.8
5 月	*	1.3	-2.2	-1.0	-0.9	-7.6
6 月	*	4.2	-1.4	-1.0	-1.8	*
7 月	*	4.0	-1.4	0.2	-2.3	*
8 月	-1.1	-0.3	-1.4	-0.3	-0.3	*
9 月	-4.6	-4.1	-0.1	-0.4	0.4	*
10 月	-3.0	-6.5	-1.7	-1.2	1.9	*
11 月	-6.4	-7.0	-1.8	1.3	4.4	*
12 月	-8.9	-4.8	-2.8	-0.8	5.9	*
1 月	*	-5.5	2.8	-1.0	1.8	0.1
2 月	*	-4.9	2.6	-1.6	2.7	-0.1
3 月	*	-5.0	3.8	-0.7	2.0	-2.5
4 月	*	-1.0	5.2	-0.7	0.5	*
5 月	*	0.6	4.1	-0.3	2.1	*
6 月	*	-0.5	1.7	-0.7	2.2	*
7 月	*	0.5	-1.3	-0.8	-0.7	*
8 月	*	-1.3	-1.4	-0.8	0.4	*
9 月	*	-2.8	-1.5	0.5	0.8	*
10 月	-5.1	-2.1	-1.5	-0.7	2.2	*
11 月	-6.2	-1.6	-0.7	-0.8	2.5	*
12 月	-6.8	1.9	-1.0	1.8	1.6	*

(b) 3-月及 7-月移动平均数据图见图 18-8, 18-9.

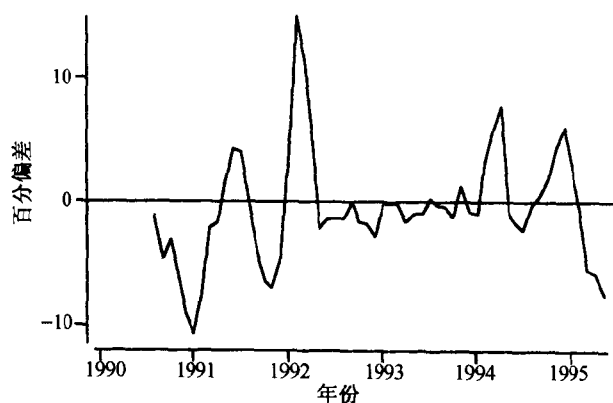


图 18-8 3-月移动平均

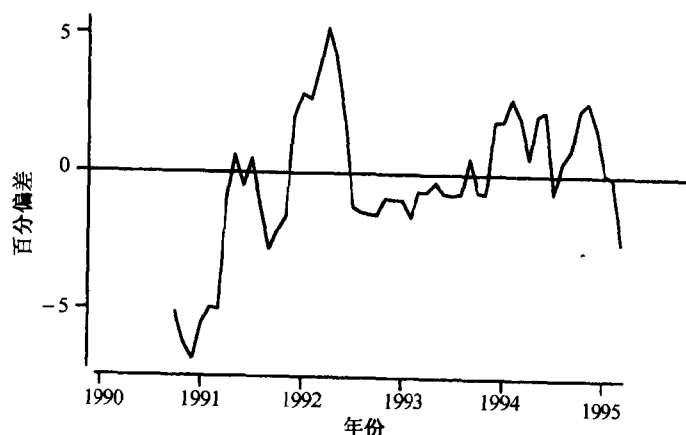


图 18-9 7-月移动平均

(c) 我们希望移动平均能消除习题 18.17 数据的不规则性, 这点可通过比较图 18-8, 18-9 与图 18-7 看出来. 从图中还可看出 7-月移动平均比 3-月移动平均图更光滑些. 3-月移动平均的大部分波动低于 10%, 而 7-月移动平均的波动低于 5%.

数据的可比性

18.20 若考虑到 1992 年为闰年, 怎样修正习题 18.9 的数据?

解 闰年的 2 月有 29 天而非 28 天. 为了便于比较, 用 $28/29$ 乘以闰年 2 月的数据. 因此在习题 18.9 的表 18.12 中, 用 $(28/29) \times 78.8 = 76.1$ 来替换 1992 年 2 月的数据 78.8. 在计算习题 18.9 到 18.13 的季节指数时, 并未对数据进行这样的调整, 但它对结论并没有多大的影响, 所以可以忽略不计.

预测

18.21 (a) 用习题 18.9 中表 18.12 的数据, 预测美国 1996 年的月新建住宅数.

(b) 比较预测值与真实值.

(c) 给出 Minitab 预测.

解 (a) 时间序列中的预测值是在已知值的基础上做出的预测(或估计). 在进行预测时, 一般仅用到趋势与季节因素, 即考虑 $Y = TS$ 而非 $Y = TSCI$. 循环及不规则因素比趋势和季节因素要难预测得多.

假设可通过寻找与表 18.12 的数据相拟合的最小二乘直线找出趋势成分. 在习题 18.13 中此

直线的方程为 $Starts = 87.747 + 0.452 \text{ time}$, 其中时间从 1 变动到 72, 若让它从 73 变动到 84, 即可得到 1996 年的趋势值, 具体见表 18.30.

表 18.30

1996	趋势值
1 月	$87.747 + 0.452 \times 73 = 120.7$
2 月	$87.747 + 0.452 \times 74 = 121.2$
3 月	$87.747 + 0.452 \times 75 = 121.6$
4 月	$87.747 + 0.452 \times 76 = 122.1$
5 月	$87.747 + 0.452 \times 77 = 122.5$
6 月	$87.747 + 0.452 \times 78 = 123.0$
7 月	$87.747 + 0.452 \times 79 = 123.4$
8 月	$87.747 + 0.452 \times 80 = 123.9$
9 月	$87.747 + 0.452 \times 81 = 124.3$
10 月	$87.747 + 0.452 \times 82 = 124.8$
11 月	$87.747 + 0.452 \times 83 = 125.2$
12 月	$87.747 + 0.452 \times 84 = 125.7$

表 18.26 给出的 Minitab 季节指数, 也可把它与趋势因素结合起来进行预测, 具体情况见表 18.31.

表 18.31

1996	趋势值(T)	季节指数(S)	预测值(TS)
1 月	120.7	0.709	85.6
2 月	121.2	0.718	87.0
3 月	121.6	0.943	114.7
4 月	122.1	1.153	140.8
5 月	122.5	1.179	144.5
6 月	123.0	1.200	147.6
7 月	123.4	1.086	134.1
8 月	123.9	1.114	138.0
9 月	124.3	1.083	134.7
10 月	124.8	1.113	138.9
11 月	125.2	0.904	113.2
12 月	125.7	0.796	100.1

(b) 1996 年新建住宅的真实数据见表 18.32.

表 18.33 展示了 1996 年每月的实际新建住宅数, 预测值以及百分数误差.

表 18.32

月份	1	2	3	4	5	6	7	8	9	10	11	12
1996 年	90.7	95.9	116.0	146.6	143.9	138.0	137.5	144.2	128.7	130.8	111.5	93.1

表 18.33

月份	1	2	3	4	5	6	7	8	9	10	11	12
1996 年	90.7	95.9	116.0	146.6	143.9	138.0	137.5	144.2	128.7	130.8	111.5	93.1
预测值	85.6	87.0	114.7	140.8	144.5	147.6	134.1	138.0	134.7	138.9	113.2	100.1
%误差	5.6	9.3	1.1	4.0	0.4	6.9	2.5	4.3	4.6	6.2	1.5	7.5

(c) 用 Minitab 下拉菜单 Stats→Time series→Decomposition, 选择预测即可得到如下输出.

Time Series Decomposition

Data Starts

Length 72.0000

NMissing 0

TrendLine Equation

$$Y_t = 87.7470 + 0.451756 * t$$

Seasonal Indices

Period	Index
1	0.708730
2	0.718379
3	0.943310
4	1.15295
5	1.17947
6	1.20045
7	1.08609
8	1.11415
9	1.08333
10	1.11307
11	0.903956
12	0.766113

Forecasts

Row	Period	Forecast
1	73	85.562
2	74	87.051
3	75	114.734
4	76	140.752
5	77	144.523
6	78	147.636
7	79	134.063
8	80	138.029
9	81	134.701
10	82	138.901
11	83	113.214
12	84	100.067

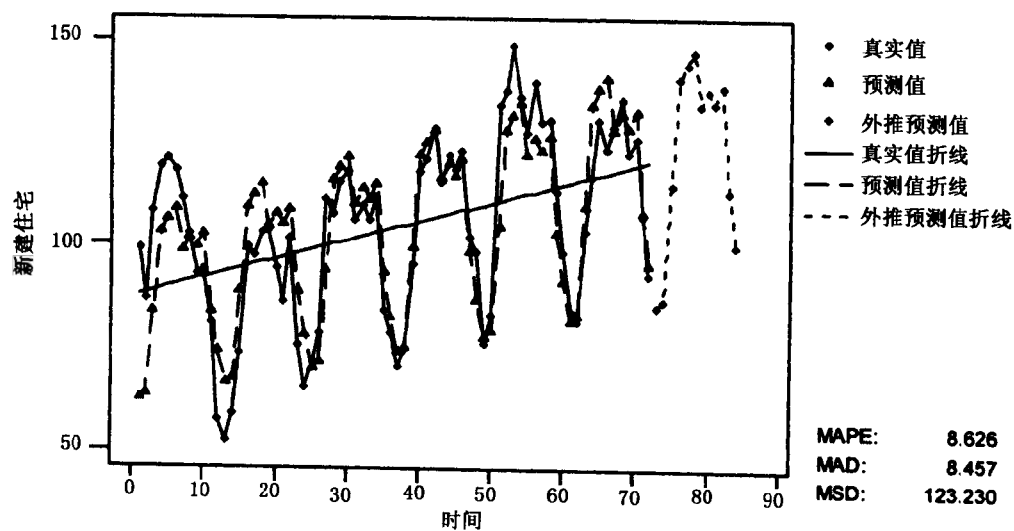


图 18-10 新建住宅的分解拟合

注意,此预测值与表 18.33 给出的预测值相同.

Minitab 也给出了图形 18-10,从中可明显看出预测值.

表 18.12 的数据图用实线表示,预测值用虚线表示,预测值的时间从 73 变动到 84.

补充习题

时间序列的特征运动

18.22 下列情形属于时间序列的哪种特征运动?

- (a) 经济衰退, (b) 夏季就业机会的增多, (c) 因科学的发展而引起死亡率的下降, (d) 钢铁工人罢工, (e) 小车需求的持续增长.

移动平均

18.23 对数字 1, 0, -1, 0, 1, 0, -1, 0, 计算

- (a) 2 阶, (b) 3 阶, (c) 4 阶, (d) 5 阶各阶移动平均.

18.24 证明:若某一序列的周期为 N (即每隔 N 项就开始重复), 则任一阶数小于 N 的移动平均序列的周期皆为 N , 并用习题 18.23 来验证.

18.25 (a) 若在习题 18.24 中进行 N 阶移动平均, 将会有什么情况发生?

- (b) 若(a)中移动平均的阶数大于 N , 结果又如何? 参考习题 18.23 来考虑.

18.26 证明:若序列的每一元素均加上(或减去)一定常数, 则移动平均序列也应增加(或减少)同一定常数.

18.27 证明:若序列的每一元素均乘以(或除以)一非零定常数, 则移动平均序列也应乘上(或除以)同一定常数.

18.28 计算习题 18.23 中(b), (c), (d)的加权移动平均, 其中权重分别为(b) 1, 2, 1, (c) 1, 2, 2 和 1, (d) 1, 2, 2, 2 和 1. 并把它与习题 18.23 的结论相比较.

18.29 (a) 在加权移动平均下, 证明习题 18.26 的性质.

- (b) 习题 18.24 的结论对加权移动平均适合吗?

18.30 一序列有(a) 24, (b) 25, (c) 200 个元素, 若对它们进行 5-阶移动平均, 则移动平均序列长度各为多少?

18.31 某序列有 M 个元素.

- (a) 证明它的 N -阶移动平均序列有 $M - N + 1$ 个元素, 对不同的 M 和 N , 通过几个例子来验证.
(b) 当 $M = N$ 时再进行讨论.

18.32 表 18.34 展示了 1986~1995 年美国的离婚及无效婚姻数目(以千计). 试构造

- (a) 2-年移动平均, (b) 2-年中心移动平均, (c) 3-年移动平均, (d) 4-年中心移动平均, (e) 6-年移动平均.

表 18.34

年份	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
离婚及无效婚姻数	1178	1166	1167	1157	1182	1189	1215	1187	1191	1169

来源:美国国家健康统计中心.

18.33 绘制习题 18.32 的原始数据图及移动平均图, 并对结论进行讨论.

18.34 (a) 验证习题 18.32(b)的 2-年中心移动平均和 3-年加权移动平均完全一致, 其中权重分别为 1, 2, 1. 再用直接数值计算方法来验证.

- (b) 证明习题 18.32(e)的 6-年中心移动平均等于某个适当权重的加权移动平均.

18.35 (a) 对习题 18.32 的数据, 计算权重分别为 1, 4, 1 的 3 阶加权移动平均.

- (b) 绘制移动平均图, 并把它与习题 18.32(c)的结果相比较.

18.36 表 18.35 展示了 1994~1996 年从巴西的月进口值(单位:百万美元). 计算(a) 12-月移动平均, (b) 12-月中心移动平均, (c) 6-月中心移动平均. 绘制移动平均图, 其上叠加原始数据图, 比较结果.

表 18.35

月份	1	2	3	4	5	6	7	8	9	10	11	12
1994	686	569	741	645	739	762	768	783	842	801	677	671
1995	805	633	745	647	702	732	715	812	692	775	775	797
1996	741	633	686	716	723	737	729	859	732	706	747	764

来源:美国调查局,美国商品贸易.

趋势估计

- 18.37** 用半平均法,计算习题 18.32 数据的趋势值,平均分别采用(a) 均值,(b) 中位数.并绘制图形.
- 18.38** 分别用(a) 手画法,(b) 适当阶的移动平均法求解习题 18.32,并把它与习题 18.37 的结果相比较.
- 18.39** (a) 用最小二乘法拟合习题 18.32 数据的直线.
(b) 从(a)的结论中找出趋势值,并把它与习题 18.37, 18.38 的结果相比较.
- 18.40** (a) 用表 18.13 的月均值,对习题 18.9 的数据拟合一条抛物线 $Y = a_0 + a_1X + a_2X^2$.
(b) 把它与习题 18.11 的最小二乘直线相比较,并计算趋势值.
- 18.41** 参考习题 18.36 的原始数据图,用(a) 半平均法,(b) 手画法,(c) 12-月中心移动平均,(d) 适当的最小二乘曲线计算习题 18.36 数据的趋势值,并讨论各种方法的优缺点.

季节变差的估计;季节指数

- 18.42** 表 18.36 展示了 1990~1995 年美国的新建家庭的月新建住宅数(以千计).
(a) 绘制数据图.
(b) 用百分数平均法计算季节指数,在计算之前,先对闰年的数据进行调整.

表 18.36

月份	1990	1991	1992	1993	1994	1995
1 月	67.9	39.2	58.4	62.8	67.2	63.6
2 月	65.9	46.1	69.2	65.5	70.8	65.3
3 月	83.2	61.4	90.9	84.9	114.6	85.3
4 月	90.0	82.8	93.5	104.4	114.3	93.9
5 月	92.4	84.5	100.2	109.2	122.3	102.3
6 月	88.9	86.8	102.7	110.1	117.6	100.5
7 月	85.5	87.4	93.2	100.4	110.4	102.0
8 月	75.6	78.7	91.8	108.3	110.1	108.5
9 月	71.9	73.7	91.4	100.6	105.2	97.7
10 月	75.6	80.9	96.1	105.5	101.3	101.5
11 月	54.9	62.6	74.8	90.6	87.8	82.0
12 月	43.1	56.3	67.9	83.3	76.8	73.7

来源:美国人口调查局,当前建筑报告.

- 18.43** 用百分数趋势法计算习题 18.42 数据的季节指数,拟合一条适当的最小二乘曲线去计算趋势值.
- 18.44** 用百分数移动平均法计算习题 18.42 数据的季节指数.
- 18.45** 用类似 Minitab 的统计软件包计算习题 18.42 数据的季节指数.
- 18.46** 比较习题 18.42 至习题 18.45 的结论.
- 18.47** 表 18.37 展示了 1990~1995 年出口到加拿大的月出口值(单位:百万美元).
(a) 绘制数据图.
(b) 用百分数平均法计算季节指数.

表 18.37

月份	1990	1991	1992	1993	1994	1995
1 月	6.3	6.8	6.9	6.9	7.6	10.1
2 月	6.7	6.4	7.0	7.7	8.2	10.2
3 月	8.0	7.1	8.2	9.5	10.4	11.7
4 月	7.4	7.6	7.8	8.8	9.4	10.6
5 月	7.9	7.7	7.7	8.8	10.0	11.4
6 月	7.5	7.5	8.4	9.1	10.2	10.9
7 月	6.2	6.5	6.9	7.1	7.6	8.4
8 月	6.7	6.8	7.0	8.3	9.9	10.8
9 月	6.4	7.4	7.9	8.6	10.2	10.8
10 月	7.5	8.3	8.0	8.9	10.5	11.4
11 月	7.4	7.0	7.7	8.9	10.6	11.1
12 月	5.9	6.1	7.1	7.9	9.8	9.7

来源:美国调查局,美国商品贸易。

- 18.48** 用趋势百分数法计算习题 18.47.
- 18.49** 用百分数移动平均法计算习题 18.47.
- 18.50** 用类似 Minitab 的统计软件包计算习题 18.47 数据的季节指数.
- 18.51** 比较习题 18.47 至习题 18.50 求得的季节指数.
- 18.52** 用两种方法求习题 18.36 的季节指数,并对结果进行比较.
- 18.53** (a) 对于习题 18.9 的数据,计算后三年和前三年的季节指数,方法任意.
(b) 对(a)的结果进行比较.
- 18.54** 若先对数据进行闰年的调整,重新计算习题 18.42 至 18.45.判断此调整是否和最后的季节指数有显著关系?

数据的消季节化

- 18.55** (a) 用习题 18.42 至 18.45 求得的任一季节指数,对习题 18.42 的数据进行消季节化.
(b) 绘制消去季节因素的数据图,并解释结论.
- 18.56** (a) 调整习题 18.47 数据的季节变差,可用习题 18.47 至 18.51 的任何结论.
(b) 绘制消去季节因素的数据图,并解释结论.
- 18.57** (a) 用习题 18.52 求得的两种季节指数,对习题 18.36 的数据进行消季节化.
(b) 绘制消去季节因素的数据图,并解释结论.

周期及不规则变差的估计

- 18.58** (a) 用任一种方法,对习题 18.55 的数据进行趋势修正.
(b) 绘制所得数据的图形.
(c) 对(a)中所得数据求 3-月及 7-月移动平均.
(d) 绘制(c)所得数据的图形,并对观测得到的变差进行解释,尤其要对可能存在的循环运动进行鉴别.
- 18.59** 若把习题 18.55 的数据换成习题 18.56 的数据,重新计算习题 18.58.
- 18.60** 若把习题 18.55 的数据换成习题 18.57 的数据,重新计算习题 18.58.
- 18.61** 表 18.38 展示了美国 1976~1995 每 1000 人的结婚率.

表 18.38

年份	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
结婚率	9.9	9.9	10.3	10.4	10.6	10.6	10.6	10.5	10.5	10.1
年份	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
结婚率	10.0	9.9	9.8	9.7	9.8	9.4	9.3	9.0	9.1	7.7

来源:美国国家健康统计中心,美国生死统计。

(a) 绘制数据图.

(b) 分析数据后, 讨论数据的循环因素是否明显?

18.62 对数据进行趋势及季节变差的调整时, 先后次序不同是否会对结果有所影响?

(a) 请进行理论上的讨论, (b) 用习题 18.42, 18.47 或 18.53 的时间序列进行验证.

18.63 (a) 用 12-月中心移动平均求解习题 18.19, 并绘制图形.

(b) 从(a)中你会得到什么结论?

18.64 (a) 对习题 18.17 及 18.18 得到的不规则变差, 写出其频数分布.

(b) 从(a)中得到的频数分布是否渐近于正态分布? 若如此, 请给出理论说明.

预测

18.65 (a) 用习题 18.42 的结论, 预测 1996 年新建家庭的月新建住宅数(以千计).

(b) 讨论可能的误差源.

(c) 对你的预测值与真实数据(见表 18.39)进行比较.

表 18.39

1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
68.9	74.2	96.9	117.9	111.6	115.0	109.1	115.6	99.3	101.0	82.6	68.8

来源: 美国人口调查局, 当前建筑报告.

18.66 (a) 用习题 18.47 的结论, 预测 1996 年出口到加拿大的月出口值(单位: 百万美元).

(b) 讨论可能的误差源.

(c) 对你的预测值与真实数据(见表 18.40)进行比较.

表 18.40

1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
10.3	11.2	11.6	11.5	11.5	11.3	9.6	10.9	11.7	12.1	12.1	10.3

来源: 美国调查局, 美国商品贸易.

18.67 用习题 18.40 提供的最小二乘抛物线, 预测习题 18.9 中 1996 的数据, 并把你的预测值与习题 18.21 的表 18.32 的真实数据进行比较.

18.68 表 18.41 展示了 1993 年到 1996 年耐存放与不耐存放货物的月零售清单(单位: 百万美元). 1995 年第一季度的值缺失. 用时间序列分析的方法估计缺失值. 若缺失值为 285, 290, 297, 计算估计的百分误差.

表 18.41

月份	1993	1994	1995	1996
1 月	246	259	X	297
2 月	251	263	Y	301
3 月	259	269	Z	303
4 月	260	271	301	305
5 月	258	273	300	304
6 月	256	274	296	300
7 月	254	270	291	300
8 月	254	276	295	303
9 月	263	287	304	313
10 月	279	304	323	334
11 月	286	311	331	338
12 月	263	286	299	309

来源: 美国调查局, 当前商业报告.

- 18.69 若统计时遗漏了习题 18.9 的表 18.12 中 1991 年 3 月、4 月及 5 月的新建住宅数,试用时间序列分析的方法进行估计。
- 18.70 若统计时遗漏了习题 18.47 的表 18.37 中 1994 年 10 月、11 月及 12 月的数值,试用时间序列分析的方法进行估计。

综合习题

- 18.71 分析表 18.42 和表 18.43 给出的时间序列.表 18.42 表示美国 1976 到 1995 年每 1000 人口中的出生率.表 18.43 表示 1991 年到 1996 年汽车商的月零售清单(单位:百万美元)。

表 18.42

年份	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
出生率	14.6	15.1	15.0	15.6	15.9	15.8	15.9	15.6	15.6	15.8
年份	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
出生率	15.6	15.7	16.0	16.4	16.7	16.3	15.9	15.5	15.0	13.4

来源:美国国家健康统计中心,美国生死统计。

表 18.43

月份	1991	1992	1993	1994	1995	1996
1 月	65.0	60.4	65.5	70.3	82.7	88.5
2 月	64.1	61.8	68.1	71.7	85.4	89.9
3 月	61.4	62.8	70.3	72.7	88.3	87.9
4 月	60.6	64.1	69.9	72.7	89.2	87.2
5 月	60.6	63.7	69.2	73.8	88.6	87.3
6 月	58.9	63.0	68.3	73.7	86.5	86.3
7 月	56.9	60.6	63.2	68.0	79.4	81.3
8 月	55.0	58.2	60.8	69.1	77.2	80.5
9 月	56.4	58.0	61.6	71.5	77.9	82.3
10 月	60.0	60.4	65.3	74.0	83.0	86.4
11 月	61.9	63.5	69.4	77.8	87.4	88.2
12 月	63.1	66.5	71.9	80.8	88.6	90.9

来源:美国调查局,当前商业报告。

- 18.72 表 18.44 展示了 1991~1996 年从美国出口到墨西哥的月出口值(单位:百万美元).分析此时间序列中的季节和循环因素。

表 18.44

月份	1991	1992	1993	1994	1995	1996
1 月	2395	3061	3193	3799	4001	4276
2 月	2364	3201	3289	3682	3672	4265
3 月	2353	3528	3755	4378	3921	4459
4 月	2759	3514	3614	3822	3383	4359
5 月	2838	3405	3504	4381	3781	4740
6 月	2861	3472	3648	4417	3704	4560
7 月	2929	3523	3180	4207	3466	4567
8 月	2849	3150	3254	4455	4187	4830
9 月	2740	3532	3392	4381	4062	4950
10 月	3225	3437	3346	4500	4313	5627
11 月	3043	3401	3956	4557	3968	5116
12 月	2921	3369	3451	4264	3835	5041

来源:美国调查局,美国商品贸易。

第十九章 过程统计控制和过程性能

对控制图的一般讨论

任何过程的变差或是由**一般原因**引起,或由**具体原因**引起.存在于原材料、机械和人员上的自然变差通常认为是一般原因引起的变差.由于过度的工具损耗,新操作人员,原材料性质的显著改变,更换承包商等等引起变差的原因是具体原因,有时也称为**指定原因**.控制图的目的之一就是找出或尽可能地消除变差的指定原因.控制图一般由**控制限**和一条**中心线**组成,如图 19-1 所示.有两个控制限,分别称为**控制上限(或 UCL)**和**控制下限(或 LCL)**.

若控制图上的一点落在控制限之外,则称此过程超出控制.除超出控制限的点外,还有其他异常形式也可表示过程超出控制,后面将会讨论这个问题.为了预测过程的行为,我们还是希望过程处于控制之中.

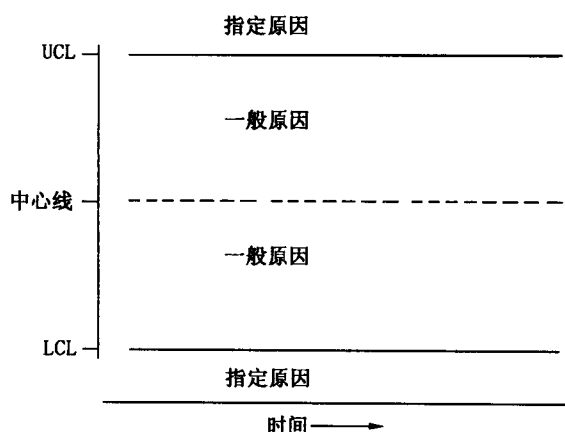


图 19-1

变量和属性控制图

控制图可分为**变量控制图**和**属性控制图**.变量及属性是和过程所收集的数据的类型相关的.当测量时间、重量、体积、长度、气压、浓度等数据时,若其连续,则认为它是**变量数据**;当考虑某产品的不合格品数时所得的数据就认为是**属性数据**.一般认为变量数据比属性数据具有优先级.表 19.1 给出各种变量和属性控制图及其统计描绘.

表 19.1

图表类型	统计图描绘
$\bar{X} - R$ 图	样本的平均值及极差图
$\bar{X} - \Sigma$ 图	样本的平均值及标准差图
中位数图	样本的中位数图
单值图	单个观测值绘图
累计和图	\bar{X} 减去目标值的累计求和图
带状图	带状权重图
EWMA 图	指数加权移动平均图
P -图	不合格品率,即不合格品数与总观测数之比
NP -图	不合格品总数
C -图	样本容量不变时,每标准单位所包含的不合格品数
U -图	样本容量变动时,每标准单位所包含的不合格品数

表 19.1 中间横线上的部分是变量控制图(或计量值控制图),下面是属性控制图(或计数值控制图).下面我们仅讨论最基本的图,且用 Minitab 构造控制图.目前借助像 Minitab 类的统计软件制图是很容易的.

$\bar{X} - R$ 图

可通过考虑均值为 μ , 标准差为 σ 的过程来了解 \bar{X} 图的思想. 假设过程由阶段性的数据组成, 每个阶段称为一个样本, 其容量为 n , \bar{X} 表示每个样本均值. 由中心极限定理知, 样本的均值是 μ , 标准差为 σ/\sqrt{n} . 样本均值的中心线在 μ 处, 控制上限与控制下限在中心线上或下 $3(\sigma/\sqrt{n})$ 处, 控制下限为

$$LCL = \mu - 3(\sigma/\sqrt{n}) \quad (1)$$

控制上限为

$$UCL = \mu + 3(\sigma/\sqrt{n}) \quad (2)$$

若过程是正态过程, 则样本均值将以 99.7% 的概率落在上下控制限之间. 实际上, 过程均值与标准差均未知, 需要估计出来. 过程均值由阶段样本均值的均值来表示, 见(3), 其中 m 是容量为 n 的阶段样本的阶段数.

$$\bar{\bar{X}} = \frac{\sum \bar{X}}{m} \quad (3)$$

均值 $\bar{\bar{X}}$ 也可通过对过程的所有数据求和, 然后再除以 mn 得到. 过程的标准差的计算方法很多, 例如合并样本方差, 或取样本标准差的均值, 或者取样本极差的均值, 或用历史值.

例 1 表 19.2 是某种产品宽度的数据. 5 个数据一组, 共有 20 组. 阶段样本组数为 $m = 20$, 每个样本容量为 $n = 5$, 所有数据之和为 199.84, 中心线为 $\bar{\bar{X}} = 1.998$. Minitab 下拉菜单 Stat→Control charts→Xbar 即可产生图 19-2. 应用下拉菜单之前, 先把表 19.2 的数据排成一列输入进去.

此过程的标准差可由四种方法估计. 用 20 个样本极差的均值, 或用 20 个样本标准差的均值, 或合并样本方差, 或用已知的历史数据. Minitab 支持这四种方法. 表 19.2 样本的 20 个均值描绘在图 19-2 上, 此图表示该过程在控制中, 各个均值随机地在中心线附近变动, 但没有一个落在控制限之外.

表 19.2

阶段 1	阶段 2	阶段 3	阶段 4	阶段 5	阶段 6	阶段 7	阶段 8	阶段 9	阶段 10
2.000	2.007	1.987	1.989	1.997	1.983	1.966	2.004	2.009	1.991
1.988	1.988	1.983	1.989	2.018	1.972	1.982	1.998	1.994	1.989
1.975	2.002	2.006	1.997	1.999	2.002	1.995	2.011	2.020	2.000
1.994	1.978	2.019	1.976	1.990	1.991	2.020	1.991	2.000	2.016
1.991	2.012	2.021	2.007	2.003	1.997	2.008	1.972	2.006	2.037
阶段 11	阶段 12	阶段 13	阶段 14	阶段 15	阶段 16	阶段 17	阶段 18	阶段 19	阶段 20
2.004	1.988	1.996	1.999	2.018	1.986	2.002	1.988	2.011	1.998
1.980	1.991	2.005	1.984	2.009	2.010	1.969	2.031	1.976	2.003
1.998	2.003	1.996	1.988	2.023	2.012	2.018	1.978	1.998	2.016
1.994	1.997	2.008	2.011	2.010	2.013	1.984	1.987	2.023	1.996
2.006	1.985	2.007	2.005	1.993	1.988	1.990	1.990	1.998	2.009

R 图用于对过程的变动进行跟踪. 对每一个样本计算极差 R , R 图的中心线由(4)给出

$$\bar{R} = \frac{\sum R}{m} \quad (4)$$

如同 \bar{X} 图, 有几种方法可估计过程的标准差.

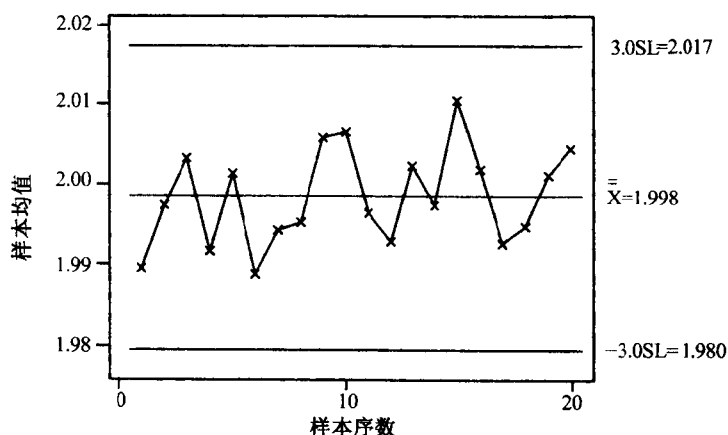


图 19-2 宽度的 \bar{X} 图

例 2 对表 19.2 的数据, 第一个样本的极差为 $R_1 = 2.000 - 1.975 = 0.025$, 第二个样本的极差为 $R_2 = 2.012 - 1.978 = 0.034$. 20 个极差分别是: 0.025, 0.034, 0.038, 0.031, 0.028, 0.030, 0.054, 0.039, 0.026, 0.048, 0.026, 0.018, 0.012, 0.027, 0.030, 0.027, 0.049, 0.053, 0.047, 0.020. 20 个极差的均值为 0.0331, 极差的 Minitab 图见图 19-3. 此 R 图并未显示出过程的变化性的任何异常行为.

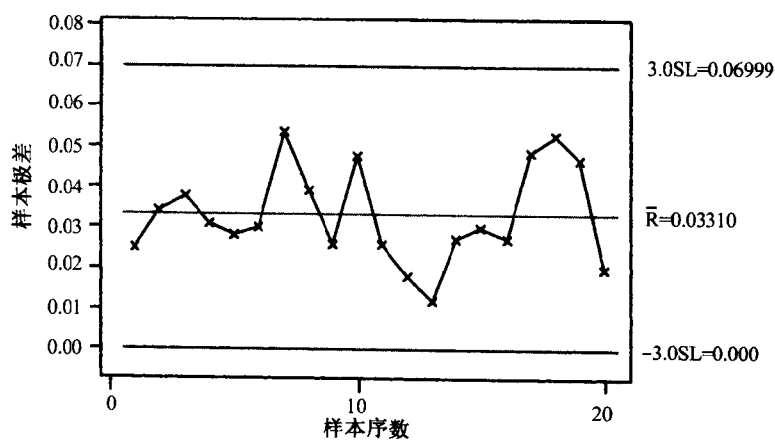


图 19-3 宽度的 R 图

指定原因的检验

指定原因可导致一点落在控制限之外, 还可造成过程的非随机化. 表 19.3 罗列了 8 种指定原因的检验方法.

表 19.3 指定原因的检验

1. 某点与中心线的距离超过 3σ .
2. 连续 9 个点位于中心线的同侧.
3. 连续 6 个点递增或递减.
4. 连续 14 个点在中心线上下交替出现.
5. 连续 3 个点中有 2 个落在距离中心线 2σ 之外(注意 3 点在中心线同侧).
6. 连续 5 个点中有 4 个点落在距离中心线 1σ 之外(注意 5 点在中心线同侧).
7. 连续 15 个点落在距离中心线 1σ 之内(两边皆可).
8. 连续 8 个点落在距离中心线 1σ 之外(两边皆可)

过程性能

一个过程只有在统计控制之下,才可进行性能分析.通常认为过程是正态的.可用 Kolmogorov-Smirnov 检验, Ryan-Joiner 检验, 或 Anderson-Darling 检验等方法来检验是否是正态过程.过程性能就是对过程行为与过程要求的比较.过程要求决定规格限, LSL 和 USL 分别表示规格下限和规格上限.

用来判断过程是否可控的数据,也可用来进行过程性能分析.均值两边 3σ 的距离称为过程散差.为研究过程统计控制而收集的数据,也可用来对过程的均值与标准差进行估计.

例 3 从例 2 知,表 19.2 的数据来自一个统计可控的过程.过程均值的估计值为 1.9984.100 个观测值的标准差为 0.013931.假定规格限为 LSL = 1.970, USL = 2.030.用 Minitab 所提供的关于正态性的 Kolmogorov-Smirnov 检验,我们并不拒绝过程是正态的.缺陷率计算如下:高于 USL 的概率为

$$P(X > 2.030) = P\left(\frac{X - 1.9984}{0.013931} > \frac{2.030 - 1.9984}{0.013931}\right) \\ = P(Z > 2.27) = 0.0116$$

即 USL 线以上有 $0.00116 \times 1000000 = 11600$ ppm(每百万单位)个是有缺陷的. $P(Z > 2.27)$ 可通过 Minitab 来计算,而不用标准正态分布表.操作如下:

MTB>cdf 2.27;

SUBC>normal 0 1.

Normal with mean = 0 and standard deviation = 1.00000

X	P(X ≤ x)
2.2700	0.9884

可见 $P(Z < 2.27) = 0.9884$, 因此 $P(Z > 2.27) = 1 - 0.9884 = 0.0116$. 低于 LSL 的概率为 $P(X < 1.970) = P(Z < -2.04) = 0.0207$. 也可用上述方法得到:

MTB>cdf -2.04;

SUBC>normal 0 1.

Normal with mean = 0 and standard deviation = 1.00000

X	P(X ≤ x)
-2.0400	0.0207

有缺陷单位的总数为 $11600 + 20700 = 32300$ ppm. 从缺陷单位来看,这个数目太大了,让人难以接受.

以 $\hat{\mu}$ 表示过程均值的估计, $\hat{\sigma}$ 表示过程标准差的估计,则缺陷率估计如下:

$$P(X > USL) = P\left(Z > \frac{USL - \hat{\mu}}{\hat{\sigma}}\right)$$

及

$$P(X < LSL) = P\left(Z < \frac{LSL - \hat{\mu}}{\hat{\sigma}}\right)$$

过程性能指数可变量过程满足规格的潜能,定义如下:

$$C_P = \frac{\text{容许散差}}{\text{测量散差}} = \frac{USL - LSL}{6\hat{\sigma}} \quad (5)$$

例 4 表 19.2 数据中, $USL - LSL = 2.030 - 1.970 = 0.060$, $6\hat{\sigma} = 6 \times 0.013931 = 0.083586$, $C_P = 0.060/0.083586 = 0.72$.

C_{PK} 指数可变量过程行为,定义如下:

$$C_{PK} = \min\left\{\frac{USL - \hat{\mu}}{3\hat{\sigma}}, \frac{\hat{\mu} - LSL}{3\hat{\sigma}}\right\} \quad (6)$$

例 5 对于例 1 的过程数据,

$$C_{PK} = \min\left\{\frac{2.030 - 1.9984}{3 \times 0.013931}, \frac{1.9984 - 1.970}{3 \times 0.013931}\right\} = \min\{0.76, 0.68\} = 0.68$$

对于仅有规格下限的过程而言,下性能指数 C_{PL} 定义如下:

$$C_{PL} = \frac{\hat{\mu} - LSL}{3\hat{\sigma}} \quad (7)$$

对于仅有规格上限的过程而言,上性能指数 C_{PU} 定义如下:

$$C_{PU} = \frac{USL - \hat{\mu}}{3\hat{\sigma}} \quad (8)$$

因此 C_{PK} 可由 C_{PL} 及 C_{PU} 定义:

$$C_{PK} = \min\{C_{PL}, C_{PU}\} \quad (9)$$

缺陷率和 C_{PL} , C_{PU} 之间具有如下关系:

$$P(X < LSL) = P\left(Z < \frac{LSL - \hat{\mu}}{\hat{\sigma}}\right) = P(Z < -3C_{PL})$$

因为

$$-3C_{PL} = \frac{LSL - \hat{\mu}}{\hat{\sigma}}$$

$$P(X > USL) = P\left(Z > \frac{USL - \hat{\mu}}{\hat{\sigma}}\right) = P(Z > 3C_{PU})$$

因为

$$3C_{PU} = \frac{USL - \hat{\mu}}{\hat{\sigma}}$$

例 6 假定 $C_{PL} = 1.1$, 则缺陷率为 $P(Z < -3 \times 1.1) = P(Z < -3.3)$, 用 Minitab 可得到计算结果.

MTB>cdf -3.3 put into c1;

SUBC>normal 0,1.

MTB>print c1;

SUBC>format (f10.8).

0.00048348

将有 $1000000 \times 0.00048348 = 483$ ppm 有缺陷的单位. 用此方法, 可构造性能指数的缺陷率表, 见表 19.4.

表 19.4

C_{PL} 或 C_{PU}	缺陷率	ppm
0.1	0.38208867	382089
0.2	0.27425308	274253
0.3	0.18406010	184060
0.4	0.11506974	115070
0.5	0.06680723	66807
0.6	0.03593027	35930
0.7	0.01786436	17864
0.8	0.00819753	8198
0.9	0.00346702	3467
1.0	0.00134997	1350
1.1	0.00048348	483
1.2	0.00015915	159
1.3	0.00004812	48
1.4	0.00001335	13
1.5	0.00000340	3
1.6	0.00000079	1
1.7	0.00000017	0
1.8	0.00000003	0
1.9	0.00000001	0
2.0	0.00000000	0

例 7 用 Minitab 对表 19.2 的数据进行性能分析,可用下拉菜单:Stat→Quality tools→Capability Analysis(Normal)实现. Minitab 输出见图 19-4. 输出给出了缺陷率,性能指数,以及一些其他的指标.例 3、例 4、例 5 的计算结果和图中相应的输出非常接近.差异因舍入误差以及不同的估计参数的方法引起.此图用直方图的形式表示样本分布,过程的总体分布用正态曲线表示. USL 右侧及 LSL 左侧的尾面积表示缺陷百分数.用 1000000 乘以百分数之和,就得到过程的 ppm 缺陷率.

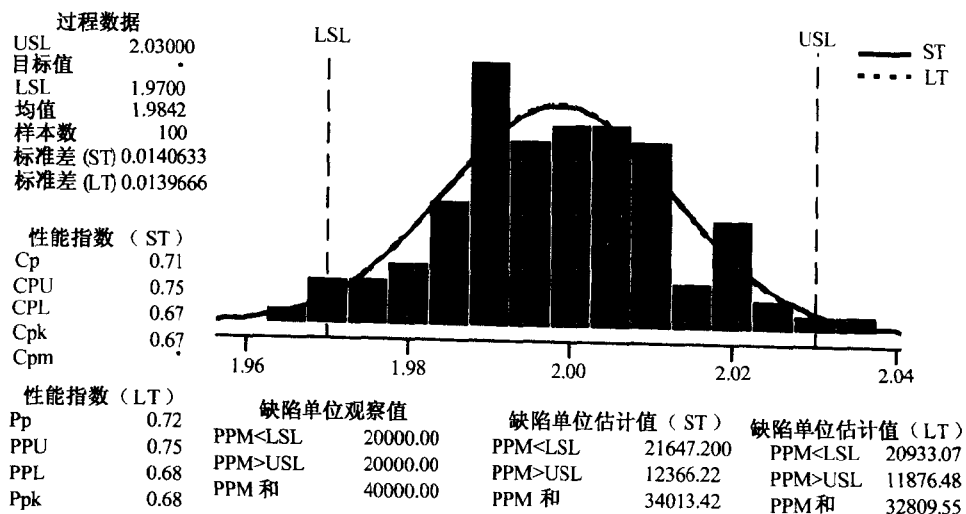


图 19-4 宽度的过程性能分析

P - 图和 NP - 图

对批量生产的产品进行分类,得到的数据称为**属性数据**.产品必须满足的标准确定后,规格就产生了.不满足此规格的产品称为**缺陷品**.不可用的缺陷品称为**不合格品**.不合格品比缺陷品所引起的后果更严重.某产品可能会因为褪色或摩擦而成为缺陷品,但它并不是不合格品.检验失误可能导致产品被划分到不合格品以及缺陷品一类.一单位产品上的疵点称为**缺陷点**,不可修复的缺陷点称为**不合格点**,不合格点比缺陷点所引起的后果更严重.

有四种控制图可用来处理属性数据,分别为:*P*-图,*NP*-图,*C*-图以及*U*-图.*P*-图和*NP*-图基于二项分布,而*C*-图和*U*-图基于泊松分布.*P*-图描绘过程的缺陷率,见例 8.

例 8 假设每隔 30 分钟对 20 个防毒面具进行检查,每 8 小时为一班组记录一次不合格品数,每一班组检查的总产品数为 $n = 20 \times 16 = 320$.表 19.5 给出 30 个班组的记录结果.*P*-图的中心线等于 30 个班组的不合格品率,即用 30 个班组检查的总产品数去除总不合格品数,或者

$$\bar{p} = 72/9600 = 0.0075$$

标准差可与二项分布联系起来,即

$$\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{0.0075 \times 0.9925}{320}} = 0.004823$$

此过程的 3σ 控制限为

$$\bar{p} \pm 3 \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad (10)$$

控制下限为 $LCL = 0.0075 - 3 \times 0.004823 = -0.006969$.当 LCL 为负值时,则认为其为零,因为不合格品率不可能小于零.控制上限为 $UCL = 0.0075 + 3 \times 0.004823 = 0.021969$.

P-图的 Minitab 解可通过下拉菜单 **Stat**→**Control charts**→**P** 得到.此例的 *P*-图见图 19-5.尽管样本 15 和 20 表明可能存在指定原因,但当它们的不合格品率两个皆等于 0.021875 和 $UCL = 0.021969$ 比较,就可看到这些点并没有超出 UCL .

表 19.5

班 组	不合格品数 X_i	不合格品率 $P_i = X/n$	班 组	不合格品数 X_i	不合格品率 $P_i = X/n$
1	1	0.003125	16	2	0.006250
2	2	0.006250	17	0	0.000000
3	2	0.006250	18	4	0.012500
4	0	0.000000	19	1	0.003125
5	4	0.012500	20	7	0.021875
6	4	0.012500	21	4	0.012500
7	4	0.012500	22	1	0.003125
8	6	0.018750	23	0	0.000000
9	4	0.012500	24	4	0.012500
10	0	0.000000	25	4	0.012500
11	0	0.000000	26	3	0.009375
12	0	0.000000	27	2	0.006250
13	1	0.003125	28	0	0.000000
14	0	0.000000	29	0	0.000000
15	7	0.021875	30	5	0.015625

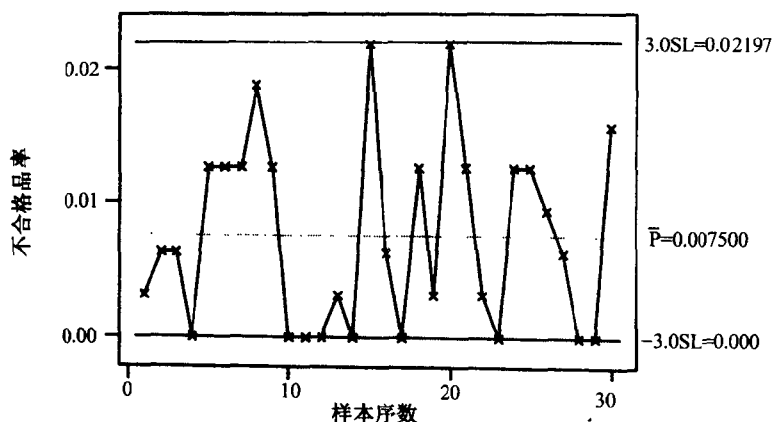


图 19-5 不合格品率的 P -图

NP -图描绘的是不合格品数而非不合格品率.许多人认为 NP -图比 P -图更合适一点,因为对于技术人员或工人来说,不合格品数要比不合格品率好统计. NP -图的中心线由 $n\bar{p}$ 给定, 3σ 控制限为

$$n\bar{p} \pm 3\sqrt{n\bar{p}(1-\bar{p})} \quad (11)$$

例 9 表 19.5 数据的中心线为 $n\bar{p} = 320 \times 0.0075 = 2.4$, 控制限为 $LCL = 2.4 - 4.63 = -2.23$, 因是负值故认为其为 0, $UCL = 2.4 + 4.63 = 7.03$. 如果某班组生产出 8 个或更多不合格品, 则此过程超出控制. 用下拉菜单 **Stat**→**Control charts**→**NP** 可得到 Minitab 解.

在执行下拉菜单之前, 必须先把每个样本的不合格品数输入到 Minitab 工作表的一列中. Minitab 的 NP -图见图 19-6.

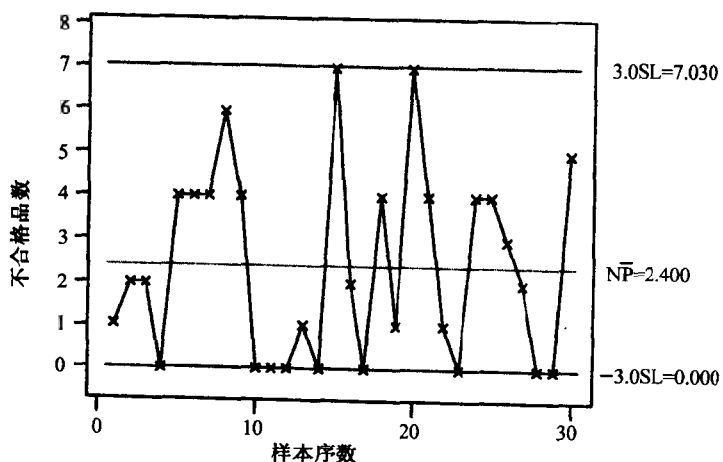


图 19-6 不合格品数的 NP -图

其他控制图

本章仅仅是对控制图应用的一个简单介绍. 表 19.1 列出了目前工业生产装置应用中的许多控制图. 为了便于在车间里计算, 有时采用**中位数图**, 即用样本中位数而非样本均值来绘图. 假如样本容量为奇数, 则样本中位数就是样本值按大小顺序排列后中间位置上的值.

若产品数较少, 则经常用到**单值控制图**. 此时样本仅包含一个观测值. 单值控制图经常被记作 X 图.

带状图分为四个带, 带 1 表示距离均值 1 个标准差以内的范围, 带 2 表示距离均值 1 个标

准差和 2 个标准差之间的范围,带 3 表示距离均值 2 个标准差和 3 个标准差之间的范围,带 4 表示距离均值 3 个标准差以外的范围.对于 4 个带分别指定权重.点落在中心线同侧的权重大一些.当各点权重的累计和大于或等于带 4 的权重时,一般就可初步认为过程超出控制.在初步认定过程超出控制后,或者当下一个点穿过中心线时,累计和就被重设为 0.

指数加权移动平均图(EWMA 图)可以作为单值图或 \bar{X} 图的备选图,它可以用来检测过程平均值与目标值之间的微小变动.EWMA 图包含了历史数据的全部信息,不仅仅是当前数据(即样本).

\bar{X} 与过程目标值的偏差的累计求和就是**累计和图**.EWMA 图和累计求和图都可用来检验过程的变动.

当我们关注的是某件产品中的单位缺陷点数或单位不合格点数而非仅仅考虑一件产品是否为合格品时,就可用 **C-图**或 **U-图**.此时有必要定义一个**观察单位**.观察单位是用来采样的输出的固定单位,采样的目的是检查其上的缺陷数.若每个样本仅取一个观察单位时,我们采用 C-控制图;当每个样本的观察单位个数不定时,采用 U-控制图进行分析.

习题及解答

$\bar{X}-R$ 图

19.1 某工序是往容器里装麦片.平均每个容器装 510 克,标准差为 5 克.每小时选择 4 个容器,他们的平均重量用来对工序进行监控,查找其指定原因,以保持工序统计可控.写出 \bar{X} 控制图的控制上限与控制下限.

解 此题中,假设 μ 和 σ 已知,分别为 510 和 5.若 μ 和 σ 未知,可通过某些方法进行估计.控制下限为 $LCL = \mu - 3(\sigma/\sqrt{n}) = 510 - 3 \times 2.5 = 502.5$,控制上限为 $UCL = \mu + 3(\sigma/\sqrt{n}) = 510 + 3 \times 2.5 = 517.5$.

19.2 表 19.6 展示了 20 个时间段上某产品宽度的数据. \bar{X} 图的控制限为 $LCL = 1.981$, $UCL = 2.081$.试判断有没有样本均值落在新的控制限之外?

表 19.6

阶段 1	阶段 2	阶段 3	阶段 4	阶段 5	阶段 6	阶段 7	阶段 8	阶段 9	阶段 10
2.000	2.007	1.987	1.989	1.997	1.983	1.966	2.004	2.009	1.991
1.988	1.988	1.983	1.989	2.018	1.972	1.982	1.998	1.994	1.989
1.975	2.002	2.006	1.997	1.999	2.002	1.995	2.011	2.020	2.000
1.994	1.978	2.019	1.976	1.990	1.991	2.020	1.991	2.000	2.016
1.991	2.012	2.021	2.007	2.003	1.997	2.008	1.972	2.006	2.037
阶段 11	阶段 12	阶段 13	阶段 14	阶段 15	阶段 16	阶段 17	阶段 18	阶段 19	阶段 20
2.004	1.988	1.996	1.999	2.018	2.025	2.002	1.988	2.011	1.998
1.980	1.991	2.005	1.984	2.009	2.022	1.969	2.031	1.976	2.003
1.998	2.003	1.996	1.988	2.023	2.035	2.018	1.978	1.998	2.016
1.994	1.997	2.008	2.011	2.010	2.013	1.984	1.987	2.023	1.996
2.006	1.985	2.007	2.005	1.993	2.020	1.990	1.990	1.998	2.009

解 20 个样本的均值分别为: 1.9896, 1.9974, 2.0032, 1.9916, 2.0014, 1.9890, 1.9942, 1.9952, 2.0058, 2.0066, 1.9964, 1.9928, 2.0024, 1.9974, 2.0106, **2.0230**, 1.9926, 1.9948, 2.0012 和 2.0044.第十六个均值, 2.0230 超出了控制上限, 其他值都在控制限之内.

- 19.3 见习题 19.2. 假定在收集第十六个样本时发生了一点意外, 因此就没有采集到第十六组数据, 此时重新计算控制限, 分别为 $LCL = 1.979$, $UCL = 2.017$. 试判断除了第十六组样本均值之外, 还有没有样本均值落在新的控制限之外?

解 除去第十六组数据, 没有其他组的均值落在新的控制限之外. 假定新图中没有出现表 19.3 中关于指定原因检验的任何情形, 则此题中给定的控制限可用来对过程进行监控.

- 19.4 验证习题 19.2 中给出的控制限. 通过合并 20 个样本方差来估计过程的标准差.

解 100 个样本观测值的均值为 1.999. 找到合并方差的方法之一是用单因素方差分析的方法去处理 20 个样本. 组内均方或误差均方就是 20 个样本的合并方差. Minitab 下的方差分析表如下:

Analysis of Variance

Source	DF	SS	MS	F	P
Factor	19	0.006342	0.000334	1.75	0.044
Error	80	0.015245	0.000191		
Total	99	0.021587			

标准差的估计值为 $\sqrt{0.000191} = 0.01382$. 控制下限为 $LCL = 1.999 - 3 \times (0.01382/\sqrt{5}) = 1.981$, 控制上限为 $UCL = 1.999 + 3 \times (0.01382/\sqrt{5}) = 2.018$.

指定原因的检验

- 19.5 表 19.7 包含 20 个样本的数据, 每个样本容量为 5. \bar{X} 图见图 19-7. 若在第 10 个时间段供应商改变, 则对此过程有什么影响? 在表 19.3 列出的指定原因的检验中, 哪种情况可能发生?

表 19.7

阶段 1	阶段 2	阶段 3	阶段 4	阶段 5	阶段 6	阶段 7	阶段 8	阶段 9	阶段 10
2.000	2.007	1.987	1.989	1.997	1.983	1.966	2.004	2.009	1.991
1.988	1.988	1.983	1.989	2.018	1.972	1.982	1.998	1.994	1.989
1.975	2.002	2.006	1.997	1.999	2.002	1.995	2.011	2.020	2.000
1.994	1.978	2.019	1.976	1.990	1.991	2.020	1.991	2.000	2.016
1.991	2.012	2.021	2.007	2.003	1.997	2.008	1.972	2.006	2.037
阶段 11	阶段 12	阶段 13	阶段 14	阶段 15	阶段 16	阶段 17	阶段 18	阶段 19	阶段 20
2.014	1.998	2.006	2.009	2.028	1.996	2.012	1.998	2.021	2.008
1.990	2.001	2.015	1.994	2.019	2.020	1.979	2.041	1.986	2.013
2.008	2.013	2.006	1.998	2.033	2.022	2.028	1.988	2.008	2.026
2.004	2.007	2.018	2.021	2.020	2.023	1.994	1.997	2.033	2.006
2.016	1.995	2.017	2.015	2.003	1.998	2.000	2.000	2.008	2.019

解 表 19.7 的控制图表明改变供应商引起了宽度的增加. 第十个时间段后的转变很明显. 图 19-7 上的数字 6 表明表 19.3 中的第六种情况发生了, 即连续 5 点中有 4 点落在距离中心线 1σ 之外 (4 点在中心线同侧). 5 点对应于 4 到 8 样本的数据.

过程性能

- 19.6 见习题 19.2. 若已知第十六个样本引起了指定原因的发生, 则可删除此样本. 通过计算其他 19 个样本的均值来估计平均宽度, 而且用他们的样本标准差来估计标准差. 假如规格限为 $LSL = 1.960$ 以及 $USL = 2.040$, 试写出下性能指数, 上性能指数以及

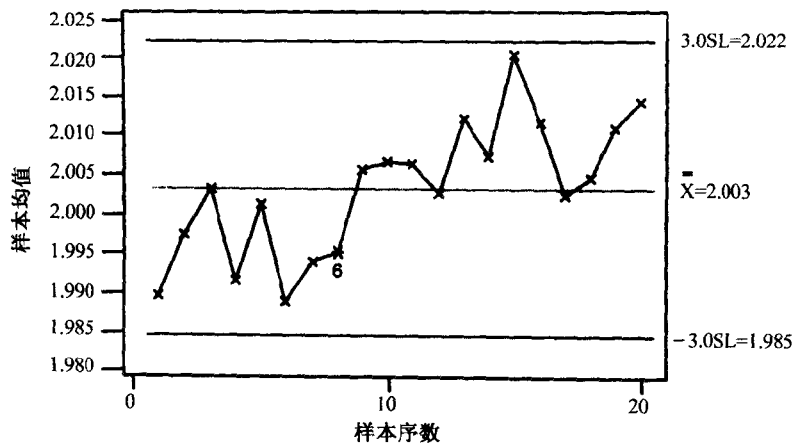


图 19-7 宽度的 \bar{X} 图

C_{PK} 指数.

解 除去第十六个样本的数据之后还剩下 95 个观测值, 计算出 $\hat{\mu} = 1.9982$, $\hat{\sigma} = 0.01400$. 则下性能指数为

$$C_{PL} = \frac{\hat{\mu} - LSL}{3\hat{\sigma}} = \frac{1.9982 - 1.960}{0.0420} = 0.910$$

上性能指数为

$$C_{PU} = \frac{USL - \hat{\mu}}{3\hat{\sigma}} = \frac{2.040 - 1.9982}{0.0420} = 0.995$$

且 $C_{PK} = \min\{C_{PL}, C_{PU}\} = 0.91$.

- 19.7** 见习题 19.1. (a) 若 $LSL = 495$, $USL = 525$, 试写出缺陷单位数; (b) 若 $LSL = 490$, $USL = 530$, 试写出缺陷单位数.

解 (a) 假设填充量服从正态分布, 则位于正态曲线下 LSL 左侧的面积计算如下:

```
MTB>cdf 495 c1;
SUBC>normal mean = 510 sigma = 5.
MTB>print c1;
SUBC>format(f10.6).
0.001350
```

由对称性知, 位于正态曲线下 USL 右侧的面积亦为 0.001350. 落在规格限外的总面积为 0.002700. 缺陷单位数 ppm 为 $0.002700 \times 1000000 = 2700$.

(b) 用同样的方法可计算出 $LSL = 490$, $USL = 530$ 时的缺陷单位数 ppm.

```
MTB>cdf 490 c1;
SUBC>normal mean = 510 sigma = 5.
MTB>print c1;
SUBC>format(f10.6).
0.000032
```

缺陷单位数 ppm 为 $0.000064 \times 1000000 = 64$.

P-图以及 NP-图

- 19.8** 检查印刷电路板以查找其中焊锡不合格者. 每天检查 500 个印刷电路板, 共检查 30 天. 每天的不合格品数见表 19.8. 构造 P-图, 查找指定原因.

表 19.8

日期	1	2	3	4	5	6	7	8	9	10
不合格品数	2	0	2	5	2	4	5	1	2	3
日期	11	12	13	14	15	16	17	18	19	20
不合格品数	3	2	0	4	3	8	10	4	4	5
日期	21	22	23	24	25	26	27	28	29	30
不合格品数	2	4	3	2	3	3	2	1	1	2

解 置信限为

$$\bar{p} \pm 3 \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

中心线为 $\bar{p} = 92/15000 = 0.00613$, 标准差为

$$\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{0.00613 \times 0.99387}{500}} = 0.00349$$

控制下限为 $0.00613 - 0.01047 = -0.00434$, 因为百分比不可能取负值, 所以认为控制下限为 0, 控制上限为 $0.00613 + 0.01047 = 0.0166$. 第十七天的不合格品率为 $P_{17} = 10/500 = 0.02$, 它是超过控制上限的惟一的不合格品率.

19.9 写出习题 19.8 数据的 NP-图的控制限.

解 不合格品数的控制限为 $n\bar{p} \pm 3 \sqrt{n\bar{p}(1-\bar{p})}$. 中心线为 $n\bar{p} = 3.067$, 则控制下限为 0, 控制上限为 8.034.

19.10 假设防毒面具一盒装 25 个或 50 个. 每隔 30 分钟在一班组内随机抽取一个盒子, 检查其中的不合格品数. 此盒中有 25 个或者 50 个防毒面具. 则每班检查的产品数从 400 变动到 800, 数据见表 19.9. 用 Minitab 找出不合格品率的控制图.

表 19.9

班组	样本容量 n_i	不合格品数 X_i	不合格品率 $P_i = X_i/n_i$
1	400	3	0.0075
2	575	7	0.0122
3	400	1	0.0025
4	800	7	0.0088
5	475	2	0.0042
6	575	0	0.0000
7	400	8	0.0200
8	625	1	0.0016
9	775	10	0.0129
10	425	8	0.0188
11	400	7	0.0175
12	400	3	0.0075
13	625	6	0.0096
14	800	5	0.0063
15	800	4	0.0050
16	800	7	0.0088
17	475	9	0.0189
18	800	9	0.0113
19	750	9	0.0120
20	475	2	0.0042

解 通过记录不合格品数来监控一个过程时,若样本容量变动,中心线还保持不变,即不合格品率是在所有样本的基础上求得的.但是标准差却随样本的变动而变动,这就提供了阶梯状的控制限.控制限为

$$\bar{p} \pm 3 \sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}}$$

中心线为 $\bar{p} = 108/11775 = 0.009172$. 对于第一个样本,已知 $n_i = 400$,

$$\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}} = \sqrt{\frac{0.009172 \times 0.990828}{400}} = 0.004767$$

以及 $3 \times 0.004767 = 0.014301$. 第一个样本的控制下限为 0, 控制上限为 $0.009172 + 0.014301 = 0.023473$. 其他班组的控制限也可同样计算. 变动的限值导致了如图 19-8 阶梯状的控制上限图.

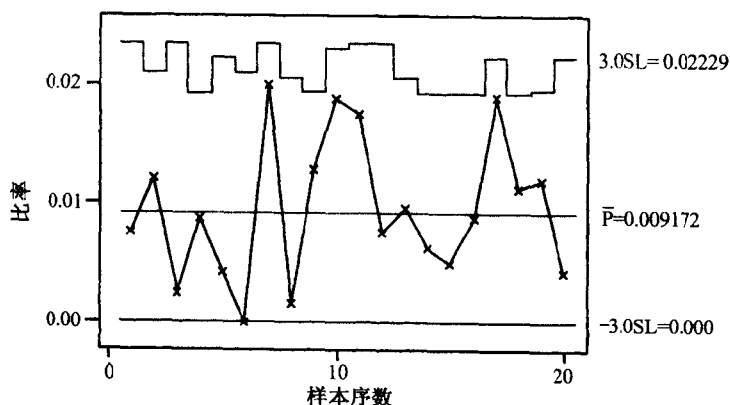


图 19-8 不合格品的 P-图

其他控制图

19.11 如果测量费用比较昂贵,或数据产出率很低,或在任何时刻输出都很均匀时,就可以采用**带移动极差的单值控制图**.数据仅由不同时刻收集的单个观测值组成.中心线为所有单个观测值的均值,变差用**移动极差**来估计.一般情况下,移动极差通过两个相邻数据差的绝对值来计算.表 19.10 给出了用于飞机的某个较昂贵的电缆的断裂强度.数据,它们是从每天的产品中选择一个电缆进行检验而得到的.给出由 Minitab 产生的单值图并解释输出结果.

表 19.10

日期	1	2	3	4	5	6	7	8	9	10
强度	491.5	502.0	505.5	499.6	504.1	501.3	503.5	504.3	498.5	508.8
日期	11	12	13	14	15	16	17	18	19	20
强度	515.4	508.0	506.0	510.9	507.6	519.1	506.9	510.9	503.9	507.4

解 用下拉菜单 Stat→Control charts→Individuals 即可得到结论.

图 19-9 就是表 19.10 数据的单值图.表 19.10 的单个观测值描绘在控制图上.控制图上第九周和第十八周的标注 2 对应于表 19.3 中的第二个指定原因的检验.因为有连续 9 个点位于中心线的同一侧,所以可能存在指定原因.第十个时间点上温度的增长导致了断裂强度的增加.断裂强度的改变导致了第十个时间段以前的点都位于中心线之下,而第十个时间段后的点大部分位于中心线之上.

19.12 指数加权移动平均图即 EWMA 图,是用来检测过程平均值与目标值 t 之间的微小变动. EWMA 图上的点由如下方程求出:

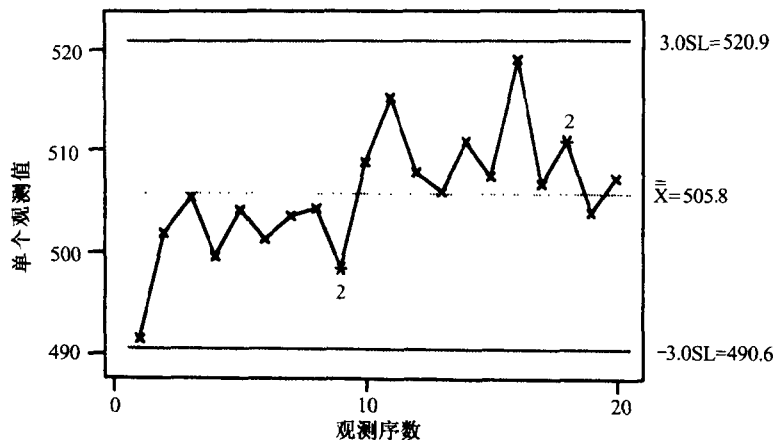


图 19-9 强度的单值图

$$\hat{x} = w\bar{x}_i + (1 - w)\hat{x}_{i-1}$$

为了展示此等式的应用,假设表 19.7 的数据从目标值为 2.000 的过程中选择而来. 初始值 \hat{x}_0 等于目标值 2.000. 权重 w 通常在 0.10 和 0.30 间变动. Minitab 中取 $w = 0.20$. 则 EWMA 图上第一个点应为 $\hat{x}_1 = w\bar{x}_1 + (1 - w)\hat{x}_0 = 0.20 \times 1.9896 + 0.80 \times 2.000 = 1.9979$, EWMA 图上第二个点应为 $\hat{x}_2 = w\bar{x}_2 + (1 - w)\hat{x}_1 = 0.20 \times 1.9974 + 0.80 \times 1.9979 = 1.9978$. 其他值同样计算. 用下拉菜单 Stat→Control charts→EWMA 可得到 Minitab 分析, 目标值确定后, Minitab 输出见图 19-10. 参考图 19-10, 分析在哪个样本上, 过程偏离了目标值?

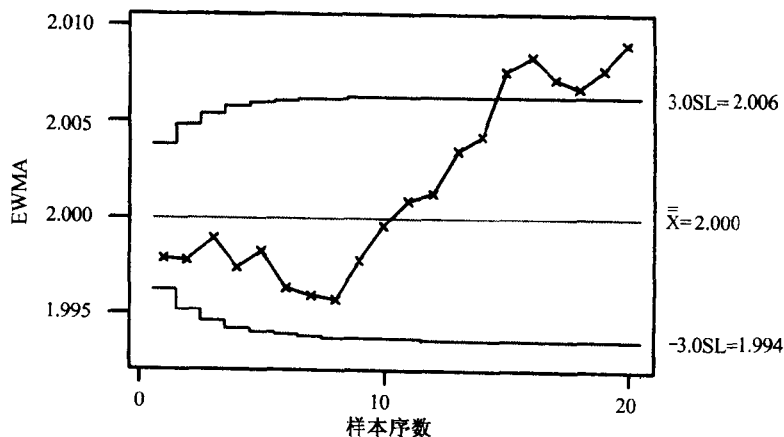


图 19-10 宽度的 EWMA 图

解 \hat{x}_i 值的图线在时间点 15 时穿过了控制上限. 因此我们得到结论, 认为过程在此点处已经偏离了目标值. 注意 EWMA 图的控制限呈阶梯状.

- 19.13** 带状图分为四个带, 带 1 表示距离均值 1 个标准差的范围, 带 2 表示距离均值 1 个标准差和 2 个标准差之间的范围, 带 3 表示距离均值 2 个标准差和 3 个标准差之间的范围, 带 4 表示距离均值 3 个标准差以外的. Minitab 中指定各个带的缺陷权重分别为 0, 2, 4, 8 (从带 1 到带 4). 点落在中心线同侧的权重大一些. 当累计和大于或等于带 4 的权重时, 一般就可初步认为过程超出控制. 在初步认定过程超出控制后, 或者当下个点穿过中心线时, 累计和就被设置为 0. 图 19-11 展示了表 19.6 数据使用带状图的 Minitab 分析. 它通过选择下拉菜单 Stat→Control charts→Zone 得到. 从此图判

断,哪个点超出了控制?

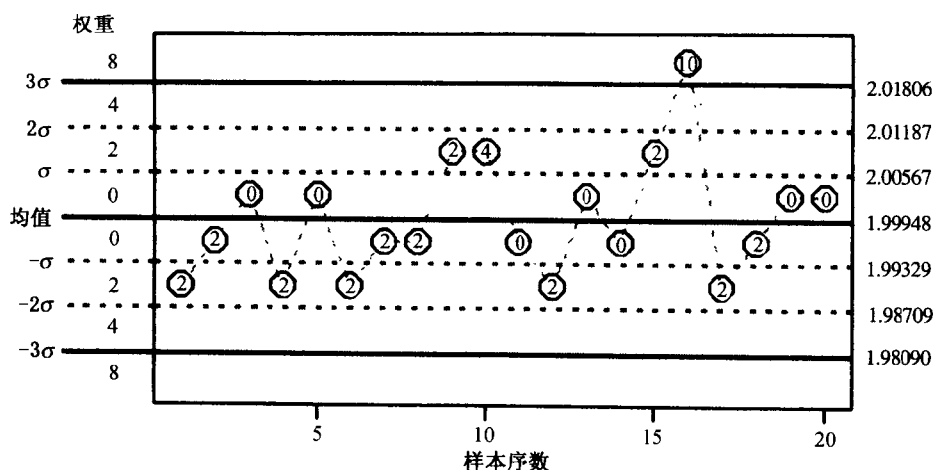


图 19-11 宽度的带状图

解 第十六组对应的点为超出控制点. 它对应的累计和为 10, 因为它超过带 4 的权重值, 就可以认为过程的此时间段为超出控制时间段.

- 19.14** 当我们关注的是某件产品的缺陷点数或不合格点数而非仅仅考虑一件产品是否为合格品时, 就可用 C -图或 U -图. 此时有必要定义一个**观察单位**. 观察单位是用来采样的输出的固定单位, 采样的目的是检查其上的缺陷点数. 若每个样本仅取一个观察单位时, 我们采用 C -控制图; 当每个样本的观察单位个数不定时, 采用 U -控制图进行分析.

C -图或 U -图可以应用在成卷产品的制造方面, 例如纸张、胶片、塑料、纤维制品等. 胶片上的黑点, 纤维制品上的污点, 针孔, 静电痕迹以及其他成卷产品上的结块, 都可认为是缺陷或疵点, 总是在生产的某个标准水平发生. C -图或 U -图就是为了确保工序的输出总是保持在缺陷发生的可接受标准范围之内. 在整卷产品中, 这些缺陷的发生一般都是随机而且相互独立的. 此时可用泊松分布形成控制图. C -图的中心线在点 \bar{c} 处, 即所有样本缺陷数的均值. 泊松分布的标准差为 $\sqrt{\bar{c}}$, 因此 3σ 控制限为 $\bar{c} \pm 3\sqrt{\bar{c}}$, 即控制下限为 $LCL = \bar{c} - 3\sqrt{\bar{c}}$, 控制上限为 $UCL = \bar{c} + 3\sqrt{\bar{c}}$.

当往材料上上涂料时, 就可能发生称为结块的缺陷. 对于某个大捆产品, 假定观测单位为 5 英尺, 则其缺陷数记录在表 19.11 中, 其中包含 24 个样本. 判断是否有点落在 3σ 控制限之外?

表 19.11

卷号	1	2	3	4	5	6	7	8	9	10	11	12
结块数	3	3	6	0	7	5	3	6	3	5	2	2
卷号	13	14	15	16	17	18	19	20	21	22	23	24
结块数	2	7	6	4	7	8	5	13	7	3	3	7

解 每卷产品结块数的均值就等于结块总数除以 24, 或 $\bar{c} = 117/24 = 4.875$. 标准差为 $\sqrt{\bar{c}} = 2.208$. 控制下限为 $LCL = 4.875 - 3 \times 2.208 = -1.749$. 因为它为负值, 故认为控制下限为 0. 控制上限为 $UCL = 4.875 + 3 \times 2.208 = 11.499$. 第二十个样本的结块数为 13, 超出了控制上限 11.499, 即第二十个样本为超出控制点.

- 19.15** 此问题追溯到习题 19.14. 在求解此问题之前最好先回顾一下习题 19.14. 表 19.12 给出了 20 个特大卷产品的数据, 它提供了卷数, 每卷的观测长度, 每卷的观测单位数(在

习题 19.14 中, 观测单位为 5 英尺), 每卷观测范围上的结块数以及每个观测单位上的结块数. U -图的中心线为 \bar{u} , 即通过第四列之和除以第三列之和得到. 标准差则随样本的改变而变化, 控制限呈阶梯状. 样本 i 的控制下限为 $LCL = \bar{u} - 3\sqrt{\bar{u}/n_i}$, 控制上限为 $UCL = \bar{u} + 3\sqrt{\bar{u}/n_i}$.

用 Minitab 构建此题的控制图, 并判断工序是否可控.

表 19.12

卷号	观测长度	观测单位数 n_i	结块数	$u_i = \text{第四列}/\text{第三列}$
1	5.0	1.0	6	6.00
2	5.0	1.0	4	4.00
3	5.0	1.0	6	6.00
4	5.0	1.0	2	2.00
5	5.0	1.0	3	3.00
6	10.0	2.0	8	4.00
7	7.5	1.5	6	4.00
8	15.0	3.0	6	2.00
9	10.0	2.0	10	5.00
10	7.5	1.5	6	4.00
11	5.0	1.0	4	4.00
12	5.0	1.0	7	7.00
13	5.0	1.0	5	5.00
14	15.0	3.0	8	2.67
15	5.0	1.0	3	3.00
16	5.0	1.0	5	5.00
17	15.0	3.0	10	3.33
18	5.0	1.0	1	1.00
19	15.0	3.0	8	2.67
20	15.0	3.0	15	5.00

解 U -图的中心线为 \bar{u} , 即通过第四列之和除以第三列之和得到. 标准差则随样本的改变而变化, 控制限呈阶梯状. 样本 i 的控制下限为 $LCL = \bar{u} - 3\sqrt{\bar{u}/n_i}$, 控制上限为 $UCL = \bar{u} + 3\sqrt{\bar{u}/n_i}$. 上述数据的中心线为 $\bar{u} = 123/33 = 3.73$. 用下拉菜单 Stat→Control charts→ U 可得到 Minitab 解. 表 19.12 的第三列与第四列数据提供了 Minitab 创建 U -控制图所需要的信息. 表 19.12 数据的 U -控制图见图 19-12. 从中可看出, 并没有点超出控制限.

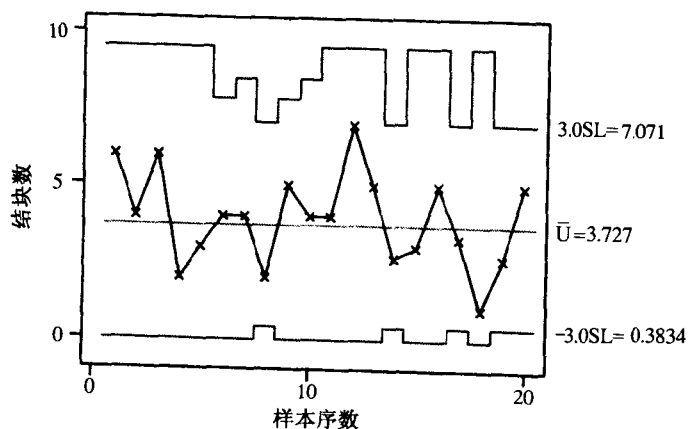


图 19-12 结块数的 U -图

补充习题

$\bar{X} - R$ 图

- 19.16 10 个容量为 4 的样本数据见表 19.13. 计算每个样本的 \bar{X} 和 R , 以及 $\bar{\bar{X}}$ 和 \bar{R} . 绘制 \bar{X} 取值的图形, 中心线为 $\bar{\bar{X}}$; 再在另一个图上绘制 R 取值的图形, 中心线为 \bar{R} .

表 19.13

样本	样本观测值			
1	13	11	13	16
2	11	12	20	15
3	16	18	20	15
4	13	15	18	12
5	12	19	11	12
6	14	10	19	16
7	12	13	20	10
8	17	17	12	14
9	15	12	16	17
10	20	13	18	17

- 19.17 某冷冻食品公司一袋包装 1 磅(454 克)绿豆. 每隔 2 小时, 选择 4 袋. 重量近似到十分位. 表 19.14 给出了一周的数据.

用习题 19.4 的方法, 通过合并 20 个样本的方差估计标准差. 用此估计计算 \bar{X} 图的控制限. 判断有没有样本均值落在控制限之外?

- 19.18 表 19.14 数据的 R 图的控制限为 $LCL=0$ 和 $UCL=8.025$. 有没有样本极差落在 3σ 控制限外?

表 19.14

周一 10:00	周一 12:00	周一 2:00	周一 4:00	周二 10:00	周二 12:00	周二 2:00	周二 4:00	周三 10:00	周三 12:00
453.0	451.6	452.0	455.4	454.8	452.6	453.6	453.2	453.0	451.6
454.5	455.0	451.5	453.0	450.9	452.8	456.1	455.8	451.4	456.0
452.6	452.8	450.8	454.3	455.0	455.5	453.9	452.0	452.5	455.0
451.8	453.5	454.8	450.6	453.6	454.8	454.8	453.5	452.1	453.0
周三 2:00	周三 4:00	周四 10:00	周四 12:00	周四 2:00	周四 4:00	周五 10:00	周五 12:00	周五 2:00	周五 4:00
454.7	451.1	452.2	454.0	455.7	455.3	454.2	451.1	455.7	450.7
451.4	452.6	448.9	452.8	451.8	452.4	452.9	453.8	455.3	452.5
450.9	448.5	455.3	455.5	451.2	452.3	451.5	452.4	455.4	454.1
455.8	454.4	453.9	453.8	452.8	452.3	455.8	454.3	453.7	454.2

- 19.19 假定为了减少袋包装的重量变动而对习题 19.17 的工序进行调整. 调整并使用一段时间之后, 得到一组新的数据, 见表 19.15. 新样本的极差由习题 19.18 所提供的控制限来描绘. 判断此调整是否使重量的变动减少? 若如此, 用表 19.15 的数据计算 \bar{X} 图的新控制限.

表 19.15

周一 10:00	周一 12:00	周一 2:00	周一 4:00	周二 10:00	周二 12:00	周二 2:00	周二 4:00	周三 10:00	周三 12:00
454.9	454.2	454.4	454.7	454.3	454.2	454.6	453.6	454.4	454.6
452.7	453.6	453.6	453.9	454.2	452.8	454.5	453.2	455.0	454.1
457.0	454.4	453.6	454.6	454.2	453.3	454.3	453.6	454.6	453.3
454.2	453.9	454.3	453.9	453.4	453.3	454.9	453.1	454.1	454.3
周三 2:00	周三 4:00	周四 10:00	周四 12:00	周四 2:00	周四 4:00	周五 10:00	周五 12:00	周五 2:00	周五 4:00
453.0	453.9	453.8	455.1	454.2	454.4	455.1	455.7	452.2	455.4
454.0	454.2	453.6	453.3	453.0	452.6	454.6	452.8	453.7	452.8
452.9	454.3	454.1	454.7	453.8	454.9	454.1	453.8	454.4	454.7
454.2	454.7	454.7	453.9	453.9	454.2	454.6	454.9	454.5	455.1

指定原因的检验

19.20 工序进行中工人频繁地调整工序会出现问题, 显然比较麻烦. 表 19.16 所包含的数据就是反映了此情形. 计算 \bar{X} 图的控制限并绘制 \bar{X} 图, 最后对表 19.3 的第八种情形进行检查.

表 19.16

阶段 1	阶段 2	阶段 3	阶段 4	阶段 5	阶段 6	阶段 7	阶段 8	阶段 9	阶段 10
2.006	2.001	1.993	1.983	2.003	1.977	1.972	1.998	2.015	1.985
1.994	1.982	1.989	1.983	2.024	1.966	1.988	1.992	2.000	1.983
1.981	1.996	2.012	1.991	2.005	1.996	2.001	2.005	2.026	1.994
2.000	1.972	2.025	1.970	1.996	1.985	2.026	1.985	2.006	2.010
1.997	2.006	2.027	2.001	2.009	1.991	2.014	1.966	2.012	2.031
阶段 11	阶段 12	阶段 13	阶段 14	阶段 15	阶段 16	阶段 17	阶段 18	阶段 19	阶段 20
2.010	1.982	2.002	1.993	2.024	1.980	2.008	1.982	2.017	1.992
1.986	1.985	2.011	1.978	2.015	2.004	1.975	2.025	1.982	1.997
2.004	1.997	2.002	1.982	2.029	2.006	2.024	1.972	2.004	2.010
2.000	1.991	2.014	2.005	2.016	2.007	1.990	1.981	2.029	1.990
2.012	1.979	2.013	1.999	1.999	1.982	1.996	1.984	2.004	2.003

过程性能

19.21 假设习题 19.17 冷冻食品包装的规格限为 $LSL = 450$ 克和 $USL = 458$ 克. 用习题 19.17 得到的 μ 和 σ 的估计值来计算 C_{PK} . 同时估计不满足规格的缺陷单位数 ppm.

19.22 若已经进行了习题 19.19 的调整, 试计算习题 19.21 的 C_{PK} , 并估计缺陷单位数 ppm.

P-图和 NP-图

19.23 某公司生产用于汽车电子系统的保险丝. 每天检查 500 个保险丝, 共检查 30 天. 表 19.17 给出了 30 天内每天检查出的不合格保险丝数. 试找出 P-图的中心线以及控制上下限, 并判断此工序是否统计可控? 若如此, 请给出不合格率的估计.

表 19.17

日期	1	2	3	4	5	6	7	8	9	10
不合格品数	3	3	3	3	1	1	1	1	6	1
日期	11	12	13	14	15	16	17	18	19	20
不合格品数	1	1	5	4	6	3	6	2	7	3
日期	21	22	23	24	25	26	27	28	29	30
不合格品数	2	3	6	1	2	3	1	4	4	5

19.24 假定在习题 19.23 中, 保险丝生产厂商想用 NP-图来代替 P-图. 试找出 NP-图的中心线以及控制上下限.

19.25 某大型百货连锁店肉类制品分部经理 Scottie Long 对有轻微褪色的牛肉包装袋的百分数特别感兴趣. 每天检测的袋数不定, 记录下有轻微褪色的袋数. 数据见表 19.18. 试给出 20 个样本的阶梯状的控制上限.

表 19.18

日期	样本容量	褪色数	褪色百分数
1	100	1	1.00
2	150	1	0.67
3	100	0	0.00
4	200	1	0.50
5	200	1	0.50
6	150	0	0.00
7	100	0	0.00
8	100	0	0.00
9	150	0	0.00
10	200	2	1.00
11	100	1	1.00
12	200	1	0.50
13	150	3	2.00
14	200	2	1.00
15	150	1	0.67
16	200	1	0.50
17	150	4	2.67
18	150	0	0.00
19	150	0	0.00
20	150	2	1.33

其他控制图

19.26 求解此问题之前先回顾一下习题 19.11. 24 小时内每隔一小时就检查一下烤箱(烤面包)的温度. 烘烤温度对工序来说是非常关键的. 数据见表 19.19. 借助单值图来监控工序的温度. 试写出中心线以及移动极差(相临两数据差的绝对值), 并考虑如何计算控制限?

表 19.19

小时	1	2	3	4	5	6	7	8	9	10	11	12
温度	350.0	350.0	349.8	350.4	349.6	350.0	349.7	349.8	349.4	349.8	350.7	350.9
小时	13	14	15	16	17	18	19	20	21	22	23	24
温度	349.8	350.3	348.8	351.6	350.0	349.7	349.8	348.6	350.5	350.3	349.1	350.0

19.27 求解此问题之前先回顾一下习题 19.12. 用 Minitab 建立表 19.14 数据的 EWMA 图. 若目标值取为 454 克, 从图中你能得到什么结论?

- 19.28 求解此问题之前先回顾一下习题 19.13 的带状图. 建立表 19.16 数据的带状图. 从中能否看出有超出控制的情形? 从此习题中, 你看出了带状图的什么缺点?
- 19.29 求解此问题之前先回顾一下习题 19.15. 建立习题 19.15 中 U -图的阶梯状的控制限.
- 19.30 Pareto 图经常用于质量控制中. 它是一个条形图, 罗列了以递减次序观察到的缺陷. 最常发生的缺陷列在第一位, 频率稍小一点的列在第二位, 依次类推. 用这样的图表可找出我们所关注的面积, 并可以试图去消除或减少缺陷百分数最大的缺陷. 下面是在给定时间段内, 检查防毒面具时所发现的缺陷: 褪色、带松、有凹痕、撕裂以及针孔. 结果见表 19.20.

表 19.20

褪色	褪色	褪色
带松	带松	带松
褪色	有凹痕	带松
褪色	带松	褪色
带松	褪色	褪色
褪色	褪色	有凹痕
褪色	有凹痕	撕裂
撕裂	针孔	褪色
有凹痕	褪色	针孔
褪色	撕裂	撕裂

图 19-13 是 Minitab 生成的 Pareto 图. 表 19.20 的数据输入到工作表的一列中. 用下拉菜单 Stat → Quality tools → Pareto charts 即可得到此图. 从此图判断, 哪种缺陷最值得关注, 其次是哪一种?

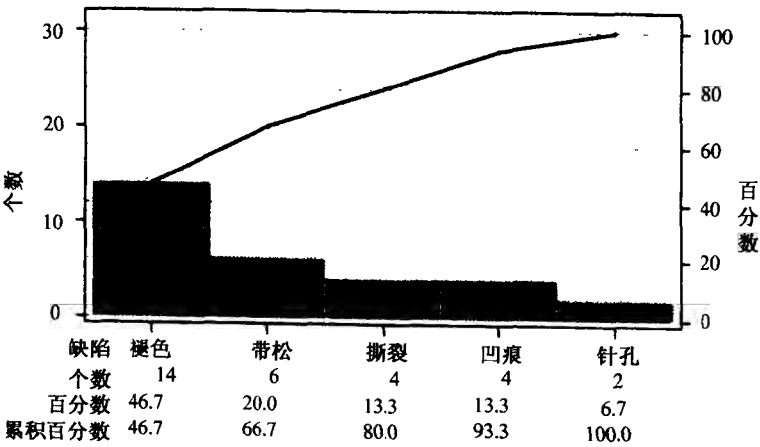


图 19-13 防毒面具缺陷图

补充习题答案

第一章

- 1.46 (a) 连续; (b) 连续; (c) 离散; (d) 离散; (e) 离散.
- 1.47 (a) 大于0; 连续. (b) 2, 3, ...; 离散. (c) 单身, 已婚, 离异, 分居, 鳏居; 离散. (d) 大于0; 连续. (e) 0, 1, 2, ...; 离散.
- 1.48 (a) 3300; (b) 5.8; (c) 0.004; (d) 46.74; (e) 126.00; (f) 4 000 000; (g) 148; (h) 0.000099; (i) 2180; (j) 43.88.
- 1.49 (a) 1 325 000; (b) 0.0041872; (c) 0.0000280; (d) 7 300 000 000; (e) 0.0003487; (f) 18.50.
- 1.50 (a) 3; (b) 4; (c) 7; (d) 3; (e) 8; (f) 无极限; (g) 3; (h) 3; (i) 4; (j) 5.
- 1.51 (a) 0.005 百万蒲式耳, 或 5000 蒲式耳; 3. (b) 0.000000005 厘米, 或 5×10^{-9} 厘米; 4. (c) 0.5 英尺; 4. (d) 0.05×10^8 米或 5×10^6 米; 2. (e) 0.5 英里/秒; 6. (f) 0.5 千英里/秒或 500 英里/秒; 3.
- 1.52 (a) 3.17×10^{-4} ; (b) 4.280×10^8 ; (c) 2.160000×10^4 ; (d) 9.810×10^{-6} ; (e) 7.32×10^5 ; (f) 1.80×10^{-3} .
- 1.53 (a) 374; (b) 14.0.
- 1.54 (a) 280(两个有效数字), 2.8 百或 2.8×10^2 ; (b) 178.9; (c) 250 000(三个有效数字), 250 千或 2.50×10^5 ; (d) 53.0; (e) 5.461; (f) 9.05; (g) 11.54; (h) 5 745 000(四个有效数字), 5745 千, 5.745 百万或 5.745×10^6 ; (i) 1.2; (j) 4157.
- 1.55 (a) -11; (b) 2; (c) $\frac{35}{8}$ 或 4.375; (d) 21; (e) 3; (f) -16; (g) $\sqrt{98}$ 或约等于 9.89961; (h) $-7/\sqrt{34}$ 或约等于 -1.20049; (i) 32; (j) $10/\sqrt{17}$ 或约等于 2.42536.
- 1.56 (a) 22, 18, 14, 10, 6, 2, -2, -6 和 -10; (b) 19.6, 16.4, 13.2, 2.8, -0.8, -4 和 -8.4; (c) -1.2, 30, $10 - 4\sqrt{2} \approx 4.34$ 和 $10 + 4\pi \approx 22.57$; (d) 3, 1, 5, 2.1, -1.5, 2.5 和 0; (e) $X = \frac{1}{4}(10 - Y)$.
- 1.57 (a) -5; (b) -24; (c) 8.
- 1.58 (a) -8; (b) 4; (c) -16.
- 1.76 (a) -4; (b) 2; (c) 5; (d) $\frac{3}{4}$; (e) 1; (f) -7.
- 1.77 (a) $a = 3, b = 4$; (b) $a = -2, b = 6$; (c) $X = -0.2, Y = -1.2$; (d) $A = \frac{184}{7} \approx 26.28571, B = \frac{110}{7} \approx 15.71429$; (e) $a = 2, b = 3, c = 5$; (f) $X = -1, Y = 3, Z = -2$; (g) $U = 0.4, V = -0.8, W = 0.3$.
- 1.78 (b) (2, -3), 即 $X = 2, Y = -3$.
- 1.79 (a) 2, -2.5; (b) 2.1 和 -0.8(约等于).
- 1.80 (a) $\frac{4 \pm \sqrt{76}}{6}$ 或约等于 2.12 和 -0.19; (b) 2 和 -2.5; (c) 约等于 0.549 和 -2.549;
(d) $\frac{-8 \pm \sqrt{-36}}{2} = \frac{-8 \pm \sqrt{36} \times \sqrt{-1}}{2} = \frac{-8 \pm 6i}{2} = -4 \pm 3i$, 其中 $i = \sqrt{-1}$.
这些根为复数, 当绘制图形时, 将不会出现.
- 1.81 (a) $-6.15 < -4.3 < -1.5 < 1.52 < 2.37$;
(b) $2.37 > 1.52 > -1.5 > -4.3 > -6.15$.
- 1.82 (a) $30 \leq N \leq 50$; (b) $S \geq 7$; (c) $-4 \leq X \leq 3$; (d) $P \leq 5$; (e) $X - Y > 2$.
- 1.83 (a) $X \geq 4$; (b) $X > 3$; (c) $N < 5$; (d) $Y \leq 1$; (e) $-8 \leq X \leq 7$; (f) $-1.8 \leq N < 3$; (g) $2 \leq a < 22$.
- 1.84 (a) 2.5877; (b) $9.5877 - 10$; (c) $8.8987 - 10$; (d) 4.1653; (e) $9.7812 - 10$; (f) $7.4464 - 10$;
(g) 2.6779; (h) 0.0030; (i) 0.8541; (j) 1.8541; (k) $6.9912 - 10$; (l) 7.9275.
- 1.85 (a) 3640; (b) 0.675; (c) 50.64; (d) 0.08445; (e) 295.1; (f) 0.0002951; (g) 0.06314; (h) 5096;
(i) 1202; (j) 2422000 或 2.422×10^6 .
- 1.86 (a) 1296000 或 1.296×10^6 ; (b) 0.05739 或 0.0574 三个有效数字; (c) 556.0; (d) 804.4; (e) 40820;
(f) 0.03438; (g) 15.51; (h) 45.67; (i) $0.0004519 = 4.519 \times 10^{-4}$ 或 4.52×10^{-4} 三个有效数字;
(j) 3096.
- 1.88 (a) $X^2 = 100 Y^3$; (b) $Y = 3 \times 10^{-2X}$.

1.89 (a) 3; (b) $\frac{3}{2}$; (c) -2; (d) -5; (e) 0.

第二章

2.19 (b) 62.

2.20 (a) 799; (b) 1000; (c) 949.5; (d) 1099.5 和 1199.5; (e) 100(小时); (f) 76; (g) $\frac{62}{400} = 0.155$ 或 15.5%; (h) 29.5%; (i) 19.0%; (j) 78.0%.

2.25 (a) 24%; (b) 11%; (c) 46%.

2.26 (a) 0.003 英寸; (b) 0.3195, 0.3225, 0.3255, ..., 0.3375 英寸;
(c) 0.320~0.322, 0.323~0.325, 0.326~0.328, ..., 0.335~0.337 英寸.

2.31 (a) 每个 5 年; (b) 4(尽管严格说来最后一组大小不明确); (c) 1; (d) (85~94); (e) 7 年和 17 年;
(f) 14.5 年和 19.5 年; (g) 49.3% 和 87.3%; (h) 45.1%; (i) 不能断定.

2.33 19.3, 19.3, 19.1, 18.6, 17.5, 19.1, 21.5, 22.5, 20.7, 18.3, 14.0, 11.4, 10.1, 18.6, 11.4 和 3.7(因为存在百分数舍入误差, 他们的和不会超过 265 百万).

2.34 (b) 0.295; (c) 0.19; (d) 0.

第三章

3.47 (a) $X_1 + X_2 + X_3 + X_4 + 8$;
(b) $f_1 X_1^2 + f_2 X_2^2 + f_3 X_3^2 + f_4 X_4^2 + f_5 X_5^2$;
(c) $U_1(U_1 + 6) + U_2(U_2 + 6) + U_3(U_3 + 6)$;
(d) $Y_1^2 + Y_2^2 + \dots + Y_N^2 - 4N$;
(e) $4X_1 Y_1 + 4X_2 Y_2 + 4X_3 Y_3 + 4X_4 Y_4$.

3.48 (a) $\sum_{j=1}^3 (X_j + 3)^3$; (b) $\sum_{j=1}^{15} f_j (Y_j - a)^2$; (c) $\sum_{j=1}^N (2X_j - 3Y_j)$;
(d) $\sum_{j=1}^8 \left(\frac{X_j}{Y_j} - 1 \right)^2$; (e) $\frac{\sum_{j=1}^{12} f_j \mu_j^2}{\sum_{j=1}^{12} f_j}$.

3.51 (a) 20; (b) -37; (c) 53; (d) 6; (e) 226; (f) -62; (g) $\frac{25}{12}$.

3.52 (a) -1; (b) 23.

3.53 86.

3.54 0.50 秒.

3.55 8.25.

3.56 (a) 82; (b) 79.

3.57 78.

3.58 66.7% 男性和 33.3% 女性.

3.59 11.09 吨.

3.60 501.0.

3.61 0.72642 厘米.

3.62 26.2.

3.63 715 分钟.

3.64 (b) 1.7349 厘米.

3.65 (a) 均值 = 5.4, 中位数 = 5; (b) 均值 = 19.91, 中位数 = 19.85.

3.66 85.

3.67 0.51 秒.

3.68 8.

3.69 11.07 吨.

3.70 490.6.

3.71 0.72638 厘米.

- 3.72 25.4.
- 3.73 约等于 78.3 年.
- 3.74 35.7 年.
- 3.75 708.3 分钟.
- 3.76 (a) 均值 = 8.9, 中位数 = 9, 众数 = 7.
(b) 均值 = 6.4, 中位数 = 6. 因为数字 4, 5, 6, 8 和 10 均出现两次, 因此我们认为他们均为众数, 不过此时认为众数不存在更合理.
- 3.77 不存在.
- 3.78 0.53 秒.
- 3.79 10.
- 3.80 11.06 吨.
- 3.81 462.
- 3.82 0.72632 厘米.
- 3.83 23.5.
- 3.84 668.7 分钟.
- 3.85 (a) 35~39; (b) 75~84.
- 3.86 (a) 利用公式(9), 众数 = 11.1; 利用公式(10), 众数 = 11.03.
(b) 利用公式(9), 众数 = 0.7264; 利用公式(10), 众数 = 0.7263.
(c) 利用公式(9), 众数 = 23.5; 利用公式(10), 众数 = 23.8.
(d) 利用公式(9), 众数 = 668.7; 利用公式(10), 众数 = 694.9.
- 3.88 (a) 8.4; (b) 4.23.
- 3.89 (a) $G = 8$; (b) $\bar{X} = 12.4$.
- 3.90 (a) 4.14; (b) 45.8.
- 3.91 (a) 11.07 吨; (b) 499.5.
- 3.92 18.9%.
- 3.93 (a) 1.01%; (b) 238.2 百万; (c) 276.9 百万.
- 3.94 \$ 1586.87.
- 3.95 \$ 1608.44.
- 3.96 3.6 和 14.4.
- 3.97 (a) 3.0; (b) 4.48.
- 3.98 (a) 3; (b) 0; (c) 0.
- 3.100 (a) 11.04; (b) 498.2.
- 3.101 38.3 英里/小时.
- 3.102 (b) 420 英里/小时.
- 3.104 (a) 25; (b) 3.55.
- 3.107 (a) 第一四分位数 = $Q_1 = 67$, 第二四分位数 = $Q_2 =$ 中位数 = 75, 第三四分位数 = $Q_3 = 83$.
(b) 分数不超过 67 的占 25%, 分数不超过 75 的占 50%, 分数不超过 83 的占 75%.
- 3.108 (a) $Q_1 = 10.55$ 吨, $Q_2 = 11.07$ 吨, $Q_3 = 11.57$ 吨; (b) $Q_1 = 469.3$, $Q_2 = 490.6$, $Q_3 = 523.3$.
- 3.109 算术平均, 中位数, 众数, Q_2 , P_{50} 和 D_5 .
- 3.110 (a) 10.15 吨; (b) 11.78 吨; (c) 10.55 吨; (d) 11.57 吨.
- 3.112 (a) 83; (b) 64.

第四章

- 4.33 (a) 9; (b) 4.273.
- 4.34 4.0 吨.
- 4.35 0.0036 厘米.
- 4.36 7.88 公斤.
- 4.37 20 周.
- 4.38 (a) 18.2; (b) 3.58; (c) 6.21; (d) 0; (e) $\sqrt{2} \approx 1.414$; (f) 1.88.

- 4.39 (a) 2; (b) 0.85.
 4.40 (a) 2.2; (b) 1.317.
 4.41 0.576 吨.
 4.42 (a) 0.00437 厘米; (b) 60.0%, 85.2% 和 96.4%.
 4.43 (a) 3.0; (b) 2.8.
 4.44 (a) 31.2; (b) 30.6.
 4.45 (a) 6.0; (b) 6.0.
 4.46 4.21 周.
 4.48 (a) 0.51 吨; (b) 27.0; (c) 12.
 4.49 3.5 周.
 4.52 (a) 1.63 吨; (b) 33.6 或 34.
 4.53 10~90 百分位极差为 \$ 189500, 销售价的 80% 落在 \$ 130250 ± \$ 94750 内.
 4.56 (a) 2.16; (b) 0.90; (c) 0.484.
 4.58 45.
 4.59 (a) 0.733 吨; (b) 38.60; (c) 12.1.
 4.61 (a) $\bar{X} = 2.47$; (b) $s = 1.11$.
 4.62 $s = 5.2$ 和极差/4 = 5.
 4.63 (a) 0.00576 厘米; (b) 72.1%, 93.3%, 99.76%.
 4.64 (a) 0.719 吨; (b) 38.24; (c) 11.8.
 4.65 (a) 0.000569 厘米; (b) 71.6%, 93.0%, 99.68%.
 4.66 (a) 146.8 磅和 12.9 磅.
 4.67 (a) 1.7349 厘米和 0.00495 厘米.
 4.74 (a) 15; (b) 12.
 4.75 (a) 统计学; (b) 代数.
 4.76 (a) 6.6%; (b) 19.0%.
 4.77 0.15.
 4.78 0.20.
 4.79 代数.
 4.80 0.19, -1.75, 1.17, 0.68, -0.29.

第五章

- 5.15 (a) 6; (b) 40; (c) 288; (d) 2188.
 5.16 (a) 0; (b) 4; (c) 0; (d) 25.86.
 5.17 (a) -1; (b) 5; (c) -91; (d) 53.
 5.19 0, 26.25, 0, 1193.1.
 5.21 7.
 5.22 (a) 0, 6, 19, 42; (b) -4, 22, -117, 560; (c) 1, 7, 38, 155.
 5.23 0, 0.2344, -0.0586, 0.0696.
 5.25 (a) $m_1 = 0$; (b) $m_2 = pq$; (c) $m_3 = pq(q - p)$; (d) $m_4 = pq(p^2 - pq + q^2)$.
 5.27 $m_1 = 0, m_2 = 5.97, m_3 = -0.397, m_4 = 89.22$.
 5.29 $m_1(\text{修正}) = 0, m_2(\text{修正}) = 5.440, m_3(\text{修正}) = -0.5920, m_4(\text{修正}) = 76.2332$.
 5.30 (a) $m_1 = 0, m_2 = 0.53743, m_3 = 0.36206, m_4 = 0.84914$;
 (b) $m_2(\text{修正}) = 0.51660, m_4(\text{修正}) = 0.78378$.
 5.31 (a) 0; (b) 52.95; (c) 92.35; (d) 7158.20; (e) 26.2; (f) 7.28; (g) 739.58; (h) 22247; (i) 706428;
 (j) 24545.
 5.32 (a) -0.2464; (b) -0.2464.
 5.33 0.9190.
 5.34 第一分布.
 5.35 (a) 0.040; (b) 0.074.

5.36 (a) -0.02 ; (b) -0.13 .

5.37

Pearson 偏度系数	分布		
	1	2	3
第一系数	0.770	0	-0.770
第二系数	1.094	0	-1.094

5.38 (a) 2.62; (b) 2.58.

5.39 (a) 2.94; (b) 2.94.

5.40 (a) 第二; (b) 第一.

5.41 (a) 第二; (b) 都不是; (c) 第一.

5.42 (a) 大于 1875; (b) 等于 1875; (c) 小于 1875.

5.43 (a) 0.313.

第六章

6.40 (a) $\frac{5}{26}$; (b) $\frac{5}{36}$; (c) 0.98; (d) $\frac{2}{9}$; (e) $\frac{7}{8}$.

6.41 (a) 第一次抽到牌 K 而第二次未抽到牌 K 的概率;
 (b) 第一次和第二次至少有一次抽到牌 K 的概率;
 (c) 第一次和第二次至少有一次未抽到牌 K 的概率;
 (d) 在第一次抽到牌 K 而第二次未抽到 K 的条件下, 第三次抽到牌 K 的概率;
 (e) 第一, 第二, 第三次均未抽到牌 K 的概率;
 (f) 第一次和第二次均抽到牌 K 或第二次未抽到牌 K 而第三次抽到牌 K 的概率.

6.42 (a) $\frac{1}{3}$; (b) $\frac{3}{5}$; (c) $\frac{11}{15}$; (d) $\frac{2}{5}$; (e) $\frac{4}{5}$.

6.43 (a) $\frac{4}{25}$; (b) $\frac{4}{75}$; (c) $\frac{16}{25}$; (d) $\frac{64}{225}$; (e) $\frac{11}{15}$; (f) $\frac{1}{5}$; (g) $\frac{104}{225}$; (h) $\frac{221}{225}$; (i) $\frac{6}{25}$; (j) $\frac{52}{225}$.

6.44 (a) $\frac{29}{185}$; (b) $\frac{2}{37}$; (c) $\frac{118}{185}$; (d) $\frac{52}{185}$; (e) $\frac{11}{15}$; (f) $\frac{1}{5}$; (g) $\frac{86}{185}$; (h) $\frac{182}{185}$; (i) $\frac{9}{37}$; (j) $\frac{26}{111}$.

6.45 (a) $\frac{5}{18}$; (b) $\frac{11}{36}$; (c) $\frac{1}{36}$.

6.46 (a) $\frac{47}{52}$; (b) $\frac{16}{221}$; (c) $\frac{15}{34}$; (d) $\frac{13}{17}$; (e) $\frac{210}{221}$; (f) $\frac{10}{13}$; (g) $\frac{40}{51}$; (h) $\frac{77}{442}$.

6.47 $\frac{5}{18}$.

6.48 (a) 81:44; (b) 21:4.

6.49 $\frac{19}{42}$.

6.50 (a) $\frac{2}{5}$; (b) $\frac{1}{5}$; (c) $\frac{4}{15}$; (d) $\frac{13}{15}$.

6.51 (a) 37.5%; (b) 93.75%; (c) 6.25%; (d) 68.75%.

6.52 (a)

X	0	1	2	3	4
$p(X)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

6.53 (a) $\frac{1}{48}$; (b) $\frac{7}{24}$; (c) $\frac{3}{4}$; (d) $\frac{1}{6}$.

6.54 (a)

X	0	1	2	3
$p(X)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{30}$

- 6.55 (a) $\frac{3}{10}$. 它表示共抽到 2 个红球的概率.
 (b) $\frac{5}{6}$. 它表示至少抽到一个红球的概率.

6.56 \$ 9.

6.57 每天 \$ 4.80.

6.58 A 拿出 \$ 12.50; B 拿出 \$ 7.50.

6.59 (a) 7; (b) 590; (c) 541; (d) 10900.

6.60 (a) 1.2; (b) 0.56; (c) $\sqrt{0.56} \approx 0.75$.

6.64 (a) 12; (b) 2520; (c) 720.

6.65 $n = 5$.

6.66 60.

6.67 (a) 5040; (b) 720; (c) 240.

6.68 (a) 8400; (b) 2520.

6.69 (a) 32805; (b) 11664.

6.70 26.

6.71 (a) 120; (b) 72; (c) 12.

6.72 (a) 35; (b) 70; (c) 45.

6.73 $n = 6$.

6.74 210.

6.75 840.

6.76 (a) 42000; (b) 7000.

6.77 (a) 120; (b) 12600.

6.78 (a) 150; (b) 45; (c) 100.

6.79 (a) 17; (b) 163.

6.81 2.95×10^{25} .

6.83 (a) $\frac{6}{5525}$; (b) $\frac{22}{425}$; (c) $\frac{169}{425}$; (d) $\frac{73}{5525}$.

6.84 $\frac{171}{1296}$.

6.85 (a) 0.59049; (b) 0.32805; (c) 0.08866.

6.86 (b) $\frac{3}{4}$; (c) $\frac{7}{8}$.

6.87 (a) 8; (b) 78; (c) 86; (d) 102; (e) 20; (f) 142.

6.90 $\frac{1}{3}$.

6.91 1/3838380.

6.92 (a) 658007 比 1; (b) 91389 比 1; (c) 9879 比 1.

6.93 (a) 649739 比 1; (b) 71192 比 1; (c) 4164 比 1; (d) 693 比 1.

6.94 $\frac{11}{36}$.

6.95 $\frac{1}{4}$.

第七章

7.35 (a) 5040; (b) 210; (c) 126; (d) 165; (e) 6.

7.36 (a) $q^7 + 7q^6p + 21q^5p^2 + 35q^4p^3 + 35q^3p^4 + 21q^2p^5 + 7qp^6 + p^7$

(b) $q^{10} + 10q^9p + 45q^8p^2 + 120q^7p^3 + 210q^6p^4 + 252q^5p^5 + 210q^4p^6 + 120q^3p^7 + 45q^2p^8 + 10qp^9 + p^{10}$

7.37 (a) $\frac{1}{64}$; (b) $\frac{3}{32}$; (c) $\frac{15}{64}$; (d) $\frac{5}{16}$; (e) $\frac{15}{64}$; (f) $\frac{3}{32}$; (g) $\frac{1}{64}$.

7.38 (a) $\frac{57}{64}$; (b) $\frac{21}{32}$.

7.39 (a) $\frac{1}{4}$; (b) $\frac{5}{16}$; (c) $\frac{11}{16}$; (d) $\frac{5}{8}$.

- 7.40 (a) 250; (b) 25; (c) 500.
- 7.41 (a) $\frac{17}{162}$; (b) $\frac{1}{324}$.
- 7.42 $\frac{64}{243}$.
- 7.43 $\frac{193}{512}$.
- 7.44 (a) $\frac{32}{243}$; (b) $\frac{192}{243}$; (c) $\frac{40}{243}$; (d) $\frac{242}{243}$.
- 7.45 (a) 42; (b) 3.550; (c) -0.1127; (d) 2.927.
- 7.47 (a) $Npq(q-p)$; (b) $Npq(1-6pq) + 3N^2p^2q^2$.
- 7.49 (a) 1.5 和 -1.6; (b) 72 和 90.
- 7.50 (a) 75.4; (b) 9.
- 7.51 (a) 0.8767; (b) 0.0786; (c) 0.2991.
- 7.52 (a) 0.0375; (b) 0.7123; (c) 0.9265; (d) 0.0154; (e) 0.7251; (f) 0.0395.
- 7.53 (a) 0.9495; (b) 0.9500; (c) 0.6826.
- 7.54 (a) 0.75; (b) -1.86; (c) 2.08; (d) 1.625 或 0.849; (e) ± 1.645 .
- 7.55 -0.995.
- 7.56 (a) 0.0317; (b) 0.3790; (c) 0.1989.
- 7.57 (a) 20; (b) 36; (c) 227; (d) 40.
- 7.58 (a) 93%; (b) 8.1%; (c) 0.47%; (d) 15%.
- 7.59 84.
- 7.60 (a) 61.7%; (b) 54.7%.
- 7.61 (a) 95.4%; (b) 23.0%; (c) 93.3%.
- 7.62 (a) 1.15; (b) 0.77.
- 7.63 (a) 0.9962; (b) 0.0687; (c) 0.0286; (d) 0.0558.
- 7.64 (a) 0.2511; (b) 0.1342.
- 7.65 (a) 0.0567; (b) 0.9198; (c) 0.6404; (d) 0.0079.
- 7.66 0.0089.
- 7.67 (a) 0.04979; (b) 0.1494; (c) 0.2241; (d) 0.2241; (e) 0.1680; (f) 0.1008.
- 7.68 (a) 0.0838; (b) 0.5976; (c) 0.4232.
- 7.69 (a) 0.05610; (b) 0.06131.
- 7.70 (a) 0.00248; (b) 0.04462; (c) 0.1607; (d) 0.1033; (e) 0.6964; (f) 0.0620.
- 7.71 (a) 0.08208; (b) 0.2052; (c) 0.2565; (d) 0.2138; (e) 0.8911; (f) 0.0142.
- 7.72 (a) $\frac{5}{3888}$; (b) $\frac{5}{324}$.
- 7.73 (a) 0.0348; (b) 0.000295.
- 7.74 $\frac{1}{16}$.
- 7.75 $p(X) = \binom{4}{X} \cdot 0.32^X \cdot 0.68^{4-X}$. 期望频数分别为 32, 60, 43, 13 和 2.
- 7.77 期望频数分别为 1.7, 5.5, 12.0, 15.9, 13.7, 7.6, 2.7 和 0.6.
- 7.78 期望频数分别为 1.1, 4.0, 11.1, 23.9, 39.5, 50.2, 49.0, 36.6, 21.1, 9.4, 3.1 和 1.0.
- 7.79 期望频数分别为 41.7, 53.4, 34.2, 14.6 和 4.7.
- 7.80 $p(X) = \frac{0.61^X e^{-0.61}}{X!}$. 期望频数分别为 108.7, 66.3, 20.2, 4.1 和 0.7.

第八章

- 8.21 (a) 9.0; (b) 4.47; (c) 9.0; (d) 3.16.
- 8.22 (a) 9.0; (b) 4.47; (c) 9.0; (d) 2.58.
- 8.23 (a) $\mu_{\bar{X}} = 22.40$ 克, $\sigma_{\bar{X}} = 0.008$ 克; (b) $\mu_{\bar{X}} = 22.40$ 克, $\sigma_{\bar{X}} =$ 稍小于 0.008 克.
- 8.24 (a) $\mu_{\bar{X}} = 22.40$ 克, $\sigma_{\bar{X}} = 0.008$ 克; (b) $\mu_{\bar{X}} = 22.40$ 克, $\sigma_{\bar{X}} = 0.0057$ 克.
- 8.25 (a) 237; (b) 2; (c) 无; (d) 34.

- 8.26 (a) 0.4972; (b) 0.1587; (c) 0.0918; (d) 0.9544.
 8.27 (a) 0.8164; (b) 0.0228; (c) 0.0038; (d) 1.0000.
 8.28 0.0026.
 8.34 (a) 0.0029; (b) 0.9596; (c) 0.1446.
 8.35 (a) 2; (b) 996; (c) 218.
 8.36 (a) 0.0179; (b) 0.8664; (c) 0.1841.
 8.37 (a) 6; (b) 9; (c) 2; (d) 12.
 8.39 (a) 19; (b) 125.
 8.40 (a) 0.0077; (b) 0.8869.
 8.41 (a) 0.0028; (b) 0.9172.
 8.42 (a) 0.2150; (b) 0.0064, 0.4504.
 8.43 0.0482.
 8.44 0.0188.
 8.45 0.0410.
 8.47 (a) 118.79 克; (b) 0.74 克.
 8.48 0.0228.
 8.49 (a) 7.2; (b) 8.4.
 8.50 (a) 106; (b) 4.
 8.51 159.
 8.52 (a) 78.7; (b) 0.0090.

第九章

- 9.21 (a) 9.5 千克; (b) 0.74 千克²; (c) 分别为 0.78 千克和 0.86 千克.
 9.22 (a) 1200 小时; (b) 105.4 小时.
 9.23 (a) 若样本容量分别为 30, 50 和 100 时, 总体标准差的估计为 101.7 小时, 101.0 小时和 100.5 小时; 各种情况下, 总体均值的估计皆为 1200 小时.
 9.24 (a) 11.09 ± 0.18 吨; (b) 11.09 ± 0.24 吨.
 9.25 (a) 0.72642 ± 0.000095 英寸; (b) 0.72642 ± 0.000085 英寸; (c) 0.72642 ± 0.000072 英寸; (d) 0.72642 ± 0.000060 英寸.
 9.26 (a) 0.72642 ± 0.000025 英寸; (b) 0.000025 英寸.
 9.27 (a) 至少 97; (b) 至少 68; (c) 至少 167; (d) 至少 225.
 9.28 (a) 至少 385; (b) 至少 271; (c) 至少 666; (d) 至少 900.
 9.29 (a) 2400 ± 45 磅, 2400 ± 59 磅; (b) 87.6%.
 9.30 (a) 0.70 ± 0.12 , 0.69 ± 0.11 ; (b) 0.70 ± 0.15 , 0.68 ± 0.15 ; (c) 0.70 ± 0.18 , 0.67 ± 0.17 .
 9.31 (a) 至少 323; (b) 至少 560; (c) 至少 756.
 9.32 (a) 16400; (b) 27100; (c) 38420; (d) 66000.
 9.33 (a) 1.07 ± 0.09 小时; (b) 1.07 ± 0.12 小时.
 9.34 (a) 0.045 ± 0.073 ; (b) 0.045 ± 0.097 ; (c) 0.045 ± 0.112 .
 9.35 (a) 63.8 ± 0.24 磅; (b) 63.8 ± 0.31 磅.
 9.36 (a) 180 ± 24.9 磅; (b) 180 ± 32.8 磅; (c) 180 ± 38.2 磅.
 9.37 8.6 磅.
 9.38 (a) 至少 4802; (b) 至少 8321; (c) 至少 11250.

第十章

- 10.29 (a) 0.2606.
 10.30 (a) 若抽取到的红球数在 22 到 42 之间, 则接受假设, 否则拒绝假设;
 (b) 0.99; (c) 若抽取到的红球数在 24 到 40 之间, 则接受假设, 否则拒绝假设.
 10.31 (a) $H_0: p = 0.5$, $H_1: p > 0.5$; (b) 单边检验;
 (c) 若抽取到的红球数大于 39, 则拒绝 H_0 , 否则接受它.

- (d) 若抽取到的红球数大于 41, 则拒绝 H_0 , 否则接受它.
- 10.32 (a) 在水平 0.05 下, 不能拒绝假设; (b) 在水平 0.05 下可以拒绝假设.
- 10.33 不管是用双边检验还是单边检验, 在水平 0.01 下都不能拒绝假设.
- 10.34 用单边检验, 在两个水平下均可拒绝此声明.
- 10.35 用单边检验, 在水平 0.05 下, 结果显著, 但在水平 0.01 下不显著.
- 10.36 是, 若均用单边检验, 在两个水平下, 结论都很显著.
- 10.37 无论采用双边还是单边检验, 在水平 0.05 下, 结果都是显著的.
- 10.38 用单边检验, 在水平 0.01 下, 结果显著. 用双边检验则得不到此结论.
- 10.39 (a) 0.3112; (b) 0.0118; (c) 0; (d) 0; (e) 0.0118.
- 10.43 (a) 8.64 ± 0.96 盎司; (b) 8.64 ± 0.83 盎司; (c) 8.64 ± 0.63 盎司.
- 10.44 控制上限分别为: (a) 6 个不合格品; (b) 4 个不合格品.
- 10.45 (a) 是; (b) 否.
- 10.46 在两个显著性水平下进行单边检验均表明: 品牌 B 比品牌 A 要好.
- 10.47 单边检验表明在水平 0.05 下, 差异显著, 但在水平 0.01 下不显著.
- 10.48 单边检验表明: 在两个水平下, 新肥料都比原来的好.
- 10.49 (a) 双边检验表明, 在水平 0.05 下, 质量间没有差异;
(b) 单边检验表明, 在水平 0.05 下, B 不比 A 好.
- 10.50 (a) 双边检验表明, 在水平 0.05 下, 不能拒绝等比例的假设;
(b) 单边检验表明, 在水平 0.05 下, A 中红球的比例比 B 要大.
- 10.51 (a) 9; (b) 10; (c) 10; (d) 8.
- 10.54 (a) 否; (b) 是; (c) 否.
- 10.55 (a) 是; (b) 是; (c) 否.
- 10.56 (a) 是; (b) 是; (c) 是.
- 10.57 (a) 否; (b) 否; (c) 否.

第十一章

- 11.20 (a) 2.60; (b) 1.75; (c) 1.34; (d) 2.95; (e) 2.13.
- 11.21 (a) 3.75; (b) 2.68; (c) 2.48; (d) 2.39; (e) 2.33.
- 11.22 (a) 1.71; (b) 2.09; (c) 4.03; (d) -0.128.
- 11.23 (a) 1.81; (b) 2.76; (c) -0.879; (d) -1.37.
- 11.24 (a) ± 4.60 ; (b) ± 3.06 ; (c) ± 2.79 ; (d) ± 2.75 ; (e) ± 2.70 .
- 11.25 (a) 7.38 ± 0.82 克; (b) 7.38 ± 1.16 克.
- 11.26 (a) 7.38 ± 0.73 克; (b) 7.38 ± 0.76 克.
- 11.27 (a) 0.298 ± 0.030 秒; (b) 0.298 ± 0.049 秒.
- 11.28 双边检验表明, 无论是在水平 0.05 还是 0.01 下, 都不能证明平均寿命已经改变.
- 11.29 单边检验表明, 无论是在水平 0.05 还是 0.01 下, 均值都未减少.
- 11.30 两个水平下的双边检验均表明, 此产品不满足所需的规格.
- 11.31 两个水平下的单边检验均表明, 铜的平均含量比所需的规格要高.
- 11.32 单边检验表明, 若显著性水平采用 0.01, 则此工序不应被引进, 但若采用 0.05, 则应被引进.
- 11.33 单边检验表明, 在显著性水平 0.05 下, 品牌 A 比品牌 B 好.
- 11.34 在显著性水平 0.05 下采用双边检验, 我们并不能从样本中得到两种型号在酸度上存在差异的结论.
- 11.35 在显著性水平 0.05 下采用单边检验, 可以认为第一组不比第二组好.
- 11.36 (a) 21.0; (b) 26.2; (c) 23.3.
- 11.37 (a) 15.5; (b) 30.1; (c) 41.3; (d) 55.8.
- 11.38 (a) 20.1; (b) 36.2; (c) 48.3; (d) 63.7.
- 11.39 (a) $\chi_1^2 = 9.59$ 和 $\chi_2^2 = 34.2$.
- 11.40 (a) 16.0; (b) 6.35; (c) 假设两个尾面积相等, 则 $\chi_1^2 = 2.17$ 和 $\chi_2^2 = 14.1$.
- 11.41 (a) 87.0 到 230.9 小时; (b) 78.1 到 288.5 小时.
- 11.42 (a) 95.6 到 170.4 小时; (b) 88.9 到 190.8 小时.

- 11.43 (a) 122.5; (b) 179.2.
 11.44 (a) 207.7; (b) 295.2.
 11.46 (a) 106.1 到 140.5 小时; (b) 102.1 到 148.1 小时.
 11.47 105.5 到 139.6 小时.
 11.48 依据给定的样本, 在任何水平下, 变异的明显增加都不显著.
 11.49 变异的明显减少在水平 0.05 下显著, 但在水平 0.01 下不显著.
 11.50 (a) $F_{0.95} = 3.07$; (b) $F_{0.99} = 4.02$; (c) $F_{0.95} = 2.11$; (d) $F_{0.99} = 2.83$.
 11.51 $F_{0.95} = 1.95$, 用内插法.
 11.52 样本 1 的变量在水平 0.05 下, 显著大些, 但在水平 0.01 下不行.
 11.53 (a) 是; (b) 否.

第十二章

- 12.26 在任何水平下, 假设都不能被拒绝.
 12.27 结论如前.
 12.28 新教师并没有采用其他人的记分方法. (事实上, 成绩比平均成绩高可能是由于教学水平或低标准两方面的原因所致)
 12.29 没有理由拒绝硬币均匀的假设.
 12.30 在任何水平下, 都没有理由拒绝假设.
 12.31 (a) 分别为 10, 60 和 50;
 (b) 在水平 0.05 下, 不能拒绝结论和所期望之相同的假设.
 12.32 在水平 0.05 下, 差异是显著的.
 12.33 (a) 拟合得好; (b) 否.
 12.34 (a) 拟合得“很好”; (b) 在水平 0.05 下, 拟合得不好.
 12.35 (a) 在水平 0.05 下, 拟合得很不好; 因为二项分布为数据提供了一个很好的拟合, 这和习题 12.33 的结论一致.
 (b) 拟合得好, 但不是“很好”.
 12.36 在水平 0.05 下, 拒绝假设, 但在水平 0.01 下不行.
 12.37 结论如前.
 12.38 在任何水平下, 都不能拒绝假设.
 12.39 在水平 0.05 下不能拒绝假设.
 12.40 在两个水平下都可以拒绝假设.
 12.41 在两个水平下都应拒绝假设.
 12.42 在两个水平下都不能拒绝假设.
 12.49 (a) 0.3863(未修正) 和 0.3779(Yates 修正).
 12.50 (a) 0.2205, 0.1985(修正); (b) 0.0872, 0.0738(修正).
 12.51 0.4651.
 12.54 (a) 0.4188, 0.4082(修正).
 12.55 (a) 0.2261, 0.2026(修正); (b) 0.0875, 0.0740(修正).
 12.56 0.3715.

第十三章

- 13.24 (a) 4; (b) 6; (c) $\frac{28}{3}$; (d) 10.5; (e) 6; (f) 9.
 13.25 (2, 1).
 13.26 (a) $2X + Y = 4$; (b) X 截距 = 2, Y 截距 = 4; (c) -2, -6.
 13.27 $Y = \frac{2}{3}X - 3$ 或 $2X - 3Y = 9$.
 13.28 (a) 斜率 = $\frac{3}{5}$, Y 截距 = -4; (b) $3X - 5Y = 11$.
 13.29 (a) $-\frac{4}{3}$; (b) $\frac{32}{3}$; (c) $4X + 3Y = 32$.

13.30 $X/3 + Y/(-5) = 1$ 或 $5X - 3Y = 15$.

13.31 (a) $^{\circ}\text{F} = \frac{9}{5}^{\circ}\text{C} + 32$; (b) 176°F ; (c) 20°C .

13.32 (a) $Y = -\frac{1}{3} + \frac{5}{7}X$, 或 $Y = -0.333 + 0.714X$;

(b) $X = 1 + \frac{9}{7}Y$, 或 $X = 1.00 + 1.29Y$.

13.33 (a) 3.24; 8.24; (b) 10.00.

13.35 (b) $Y = 29.13 + 0.661X$; (c) $X = -14.39 + 1.15Y$; (d) 79; (e) 95.

13.36 (b) 出生率 = $16.6 - 0.357 \times \text{年号}$;

(c)

年	年号	出生率	拟合值	残差
1990	0	16.6	16.6143	-0.0142
1991	1	16.3	16.2571	0.0428
1992	2	15.9	15.9000	0.0000
1993	3	15.5	15.5429	-0.0428
1994	4	15.2	15.1857	0.0142
1995	5	14.8	14.8286	-0.0285
1996	6	14.5	14.4714	0.0285

(d) 出生率预测值 = 13.0.

13.37 (b) 千数 = $2604 + 102 \times \text{年号}$.

(c)

年	年号	千数	拟合值	残差
1985	0	2667	2604.41	62.5897
1986	1	2742	2706.40	35.6037
1987	2	2823	2808.38	14.6177
1988	3	2885	2910.37	-25.3683
1989	4	2968	3012.35	-44.3543
1990	5	3022	3114.34	-92.3403
1991	6	3185	3216.33	-31.3263
1992	7	3306	3318.31	-12.3124
1993	8	3431	3420.30	10.7016
1994	9	3541	3522.28	18.7156
1995	10	3652	3624.27	27.7296
1996	11	3762	3726.26	35.7436

(d) 不小于 85 的年号的预测值 = 4644000

13.38 $Y = 5.51 + 3.20(X - 3) + 0.733(X - 3)^2$ 或 $Y = 2.51 - 1.20X + 0.733X^2$.

13.39 (b) $D = 41.77 - 1.096V + 0.08786V^2$; (c) 170 英尺, 516 英尺.

13.40 (b) 差 = $-2.68 + 1.39 \times \text{年号}$.

(c)

年	年号	男性	女性	差	拟合值	残差
1920	0	53.90	51.81	-2.09	-2.68	0.59
1930	1	62.14	60.64	-1.50	-1.28	-0.22
1940	2	66.06	65.61	-0.45	0.11	-0.56
1950	3	75.19	76.14	0.95	1.51	-0.56
1960	4	88.33	90.99	2.66	2.90	-0.24
1970	5	98.93	104.31	5.38	4.29	1.09
1980	6	110.05	116.49	6.44	5.69	0.75
1990	7	121.24	127.47	6.23	7.08	-0.85

(d) 1995 年差的预测值为 $-2.68 + 1.39 \times 7.5 = 7.75$. 趋势好象不连续.

13.41 (b) 比值 $= 0.965 + 0.0148 \times \text{年号}$.

(c)

年	年号	男性	女性	比值	拟合值	残差
1920	0	53.90	51.81	0.96	0.97	-0.00
1930	1	62.14	60.64	0.98	0.98	-0.00
1940	2	66.06	65.61	0.99	0.99	-0.00
1950	3	75.19	76.14	1.01	1.01	0.00
1960	4	88.33	90.99	1.03	1.02	0.01
1970	5	98.93	104.31	1.05	1.04	0.01
1980	6	110.05	116.49	1.06	1.05	0.00
1990	7	121.24	127.47	1.05	1.07	-0.02

(d) 比值预测值 $= 1.08$. 真实比值 $= 1.04$.

13.42 (b) 差 $= -2.63 + 1.35x + 0.0064x^2$

(d) 1995 年的预测差为 $-2.63 + 1.35 \times 7.5 + 0.0064 \times 56.25 = 7.86$.

13.43 (b) $Y = 32.14 \times 1.427^X$ 或 $Y = 32.14 \times 10^{0.1544X}$ 或 $Y = 32.14 \times e^{0.3556X}$, 其中 $e = 2.718\cdots$ (d) 387.

第十四章

14.40 (b) $Y = 4.000 + 0.500X$; (c) $X = 2.408 + 0.612Y$.

14.41 (a) 1.304; (b) 1.443.

14.42 (a) 24.50; (b) 17.00; (c) 7.50.

14.43 0.5533.

14.45 1.5.

14.46 (a) 0.8961; (b) $Y = 80.78 + 1.138X$; (c) 132.

14.47 (a) 0.958; (b) 0.872.

14.48 (a) $Y = 0.8X + 12$; (b) $X = 0.45Y + 1$.

14.49 (a) 1.60; (b) 1.20.

14.50 ± 0.80 .

14.51 75%.

14.53 (a) -0.9203.

14.54 (a) $Y = 18.04 - 1.34X$, $Y = 51.18 - 2.01X$.

14.58 0.5440.

14.59 (a) $Y = 4.44X - 142.22$; (b) 分别为 141.9 磅和 177.5 磅.

14.60 (a) 16.92 磅; (b) 2.07 英寸.

14.62 0.946.

14.63 0.269.

14.64 (a) 是; (b) 否.

14.65 (a) 否; (b) 是.

14.66 (a) 0.2923 和 0.7951; (b) 0.1763 和 0.8361.

14.67 (a) 0.3912 和 0.7500; (b) 0.3146 和 0.7861.

14.68 (a) 0.7096 和 0.9653; (b) 0.4961 和 0.7235.

14.69 (a) 是; (b) 否.

14.70 (a) 2.00 ± 0.21 ; (b) 2.00 ± 0.28 .

14.71 (a) 用单边检验, 拒绝假设;

(b) 用单边检验, 不拒绝假设.

14.72 (a) 37.0 ± 3.28 ; (b) 37.0 ± 4.45 .

14.73 (a) 37.0 ± 0.69 ; (b) 37.0 ± 0.94 .

14.74 (a) 1.138 ± 0.398 ; (b) 132.0 ± 16.6 ; (c) 132.0 ± 5.4 .

第十五章

15.26 (a) $X_3 = b_{3.12} + b_{31.2}X_1 + b_{32.1}X_2$;

(b) $X_4 = b_{4.1235} + b_{41.235}X_1 + b_{42.135}X_2 + b_{43.125}X_3$.

15.28 (a) $X_3 = 61.40 - 3.65X_1 + 2.54X_2$; (b) 40.

15.29 (a) $X_3 - 74 = 4.36(X_1 - 6.8) + 4.04(X_2 - 7.0)$ 或 $X_3 = 16.07 + 4.36X_1 + 4.04X_2$; (b) 84 和 66.

15.31 3.12.

15.32 (a) 5.883; (b) 0.6882.

15.33 0.9927.

15.34 (a) 0.7567; (b) 0.7255; (c) 0.6810.

15.37 (a) 0.5950; (b) -0.8995 ; (c) 0.8727.

15.38 (a) 0.2672; (b) 0.5099; (c) 0.4026.

15.42 (a) $X_4 = 6X_1 + 3X_2 - 4X_3 - 100$; (b) 54.

15.43 (a) 0.8710; (b) 0.8587; (c) -0.8426 .

15.44 (a) 0.8947; (b) 2.680.

第十六章

16.21 两个水平下,产量间都存在显著差异.

16.22 两个水平下,轮胎间都不存在显著差异.

16.23 在水平 0.05 下,教学方法间存在显著差异,但在水平 0.01 下不存在.

16.24 在水平 0.05 下,品牌间存在显著差异,但在水平 0.01 下不存在.

16.25 两个水平下,成绩间都存在显著差异.

16.26 工人间或机器间都不存在显著差异.

16.27 答案与习题 16.26 相同.

16.28 在水平 0.05 下,由于玉米品种不同而存在显著差异,但土壤不同不会引起显著差异.

16.29 在水平 0.01 下,玉米品种与土壤的不同均不会引起显著差异.

16.30 在水平 0.05 下,轮胎和汽车的不同均会引起显著差异.

16.31 在水平 0.01 下,轮胎和汽车的不同均不会引起显著差异.

16.32 在水平 0.05 下,教学方法间存在显著差异,但学校的不同不会引起显著差异.

16.33 无论头发颜色还是身高,都不会引起显著差异.

16.34 答案与习题 16.33 相同.

16.35 在水平 0.05 下,地区的不同会引起显著差异,但肥料的不同不会引起显著差异.

16.36 在水平 0.01 下,地区与肥料的不同都不会引起显著差异.

16.37 工人间存在显著差异,但机器间不存在.

16.38 肥料或土壤都不存在显著差异.

16.39 答案同习题 16.38.

16.40 身高,头发颜色或出生地的不同均不会引起成绩上的显著差异.

16.41 雏鸡品种及第一种化学药品的数量均会引起显著差异,但第二种化学药品或雏鸡的原始体重不会引起显著差异.

16.42 不同的电缆可能引起电缆强度上的显著差异,但不同的工人,机器或公司不会引起电缆强度的显著差异.

16.43 任何水平下都不存在显著差异.

16.44 任何水平下都不存在显著差异.

16.46 在水平 0.05 下,工人的熟练程度与其智商均可能引起测验成绩的显著差异.

16.47 在水平 0.01 下,工人的熟练程度对测验成绩的影响不显著,但智商对测验成绩的影响却很显著.

16.48 学生所处的地区的不同对他们的成绩的影响不显著,但智商对测验成绩的影响却很显著.

16.49 答案同习题 16.48.

16.53 在水平 0.05 下,化学药品和地区的不同都会引起显著差异.

- 16.54 在水平 0.05 下,地区的不同会引起显著差异,但肥料的不同不会引起显著差异.
 16.55 在水平 0.01 下,地区和肥料的不同均不会引起显著差异.
 16.56 因子 1,因子 2 或处理 A,B 和 C 都不会引起显著差异.
 16.58 因子和处理的的不同均不会引起显著差异.

第十七章

- 17.26 水平 0.05 下存在显著差异,但水平 0.01 下不存在.
 17.27 是.
 17.28 在水平 0.05 下此课程设计有效.
 17.29 在水平 0.05 下,可以拒绝销售额增长的假设.
 17.30 否.
 17.31 (a) 拒绝;(b) 接受;(c) 接受;(d) 拒绝.
 17.34 在水平 0.05 下没有差异.
 17.35 否.
 17.36 (a) 是;(b) 是.
 17.37 是.
 17.38 (a) 是;(b) 是.
 17.41 3.
 17.42 6.
 17.49 任何水平下都不存在显著差异.
 17.50 在水平 0.05 下差异显著,但在水平 0.01 下不显著.
 17.51 在水平 0.05 下差异显著,但在水平 0.01 下不显著.
 17.52 两个水平下,成绩间都存在显著差异.
 17.55 (a) 8;(b) 10.
 17.56 (a) 10;(b) 水平 0.05 下反映是随机的.
 17.62 在水平 0.05 下样本不随机.存在过多的游程,可能存在循环因素.
 17.63 在水平 0.05 下样本不随机.存在过少的游程,可能存在趋势因素.
 17.64 在水平 0.05 下,数字序列是随机的.
 17.65 (a) 在水平 0.05 下,数字序列是随机的;
 (b) 在水平 0.05 下,数字序列是随机的.
 17.69 (a) 0.67;(b) 两个裁判的判断不是很一致.

第十八章

- 18.22 (a) 循环;(b) 季节;(c) 长期;(d) 不规则;(e) 长期.
 18.23 (a) 0.5, -0.5, -0.5, 0.5, 0.5, -0.5, -0.5, 0.5;(b) $0, -\frac{1}{3}, 0, \frac{1}{3}, 0, -\frac{1}{3}, 0$;(c) 0, 0, 0, 0, 0, 0;
 (d) $\frac{1}{5}, 0, -\frac{1}{5}, 0, \frac{1}{5}$.
 18.28 (b) 0, -0.5, 0, 0.5, 0, -0.5, 0;(c) $-\frac{1}{6}, -\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, -\frac{1}{6}, -\frac{1}{6}$;
 (d) 0, 0, 0, 0, 0.
 18.30 (a) 20;(b) 21;(c) 196.

第十九章

- 19.16 样本均值为:13.25 14.50 17.25 14.50 13.50 14.75 13.75 15.00 15.00 17.00
 样本极差为:5 9 5 6 8 9 10 5 5 7
 $\bar{X} = 14.85, \bar{R} = 6.9$.
 19.17 σ 的估计值为 1.741. LCL = 450.7, UCL = 455.9. 没有样本均值落在控制限之外.
 19.18 否.
 19.19 图表明变动已经减少.新的控制限为 LCL = 452.0 UCL = 455.9. 它表明经过调整后,过程已处于受控状态.

在目标值附近.

- 19.20 控制限为 $LCL = 1.980$, $UCL = 2.017$. 第 4, 5, 6 阶段的值不能通过检验 5. 第 15 到 20 阶段的值不能通过检验 4. 它们中的每一个都可看作是上下连续波动的 14 个点的结束点.

19.21 $C_{PK} = 0.63$. 单位缺陷数 $\text{ppm} = 32487$.

19.22 $C_{PK} = 1.72$. 单位缺陷数 $\text{ppm} < 1$.

19.23 中心线 = 0.006133, $LCL = 0$, $UCL = 0.01661$; 过程可控; $\text{ppm} = 6133$.

19.24 中心线 = 3.067, $LCL = 0$, $UCL = 8.304$.

19.25 0.032 0.027 0.032 0.024 0.024 0.027 0.032 0.032 0.027 0.024

0.032 0.024 0.027 0.024 0.027 0.024 0.027 0.027 0.027 0.027

19.26 中心线 = $\bar{X} = 349.9$.

移动极差为: 0.0 0.2 0.6 0.8 0.4 0.3 0.1 0.4 0.4 0.9 0.2 1.1 0.5 1.5 2.8

1.6 0.3 0.1 1.2 1.9 0.2 1.2 0.9

上述移动极差的均值 = $\bar{R}_M = 0.765$.

单值图控制限: $\bar{X} \pm 3 \frac{\bar{R}_M}{d_2}$, 其中 d_2 是控制图常数, 可通过各种表查到, 此题中为 1.128,

$LCL = 347.9$, $UCL = 352.0$.

- 19.27 EWMA 图表明过程均值都低于目标值. 样本 12 和 13 的均值落在控制下限之外. 13 以上的样本均值都在控制下限之上, 但是, 过程均值还是始终低于目标值.

19.28 带状图并没有指出有点超出控制. 但是, 如习题 19.20 所示, 有连续 14 个点围绕中心线上下波动. 这是由带状图本身的原因引起的.

19.29 20 个控制下限为: 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.38 0.00 0.00 0.00 0.00
0.00 0.38 0.00 0.00 0.38 0.00 0.38 0.38.

20 个控制上限为: 9.52 9.52 9.52 9.52 9.52 9.52 7.82 8.46 7.07 7.82 8.46 9.52 9.52
9.52 7.07 9.52 9.52 7.07 9.52 7.07 7.07.

19.30 褪色; 褪色和带松.

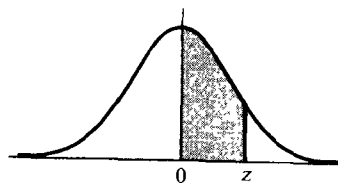
附录 I



z	0	1	2	3	4	5	6	7	8	9
0.0	.3989	.3989	.3989	.3988	.3986	.3984	.3982	.3980	.3977	.3973
0.1	.3970	.3965	.3961	.3956	.3951	.3945	.3939	.3932	.3925	.3918
0.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825
0.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697
0.4	.3683	.3668	.3653	.3637	.3621	.3605	.3589	.3572	.3555	.3538
0.5	.3521	.3503	.3485	.3467	.3448	.3429	.3410	.3391	.3372	.3352
0.6	.3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144
0.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920
0.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685
0.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1895	.1872	.1849	.1826	.1804	.1781	.1758	.1736
1.3	.1714	.1691	.1669	.1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804
1.8	.0790	.0775	.0761	.0748	.0734	.0721	.0707	.0694	.0681	.0669
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449
2.1	.0440	.0431	.0422	.0413	.0404	.0396	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.0310	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0229
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180
2.5	.0175	.0171	.0167	.0163	.0158	.0154	.0151	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0088	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.0051	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0025
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009
3.5	.0009	.0008	.0008	.0008	.0008	.0007	.0007	.0007	.0007	.0006
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003	.0003
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001

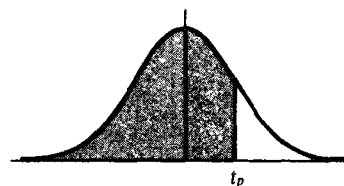
附录 II

标准正态分布的随机变量
落在 0 到 z 区间上的概率值

[illegible]

附录Ⅲ

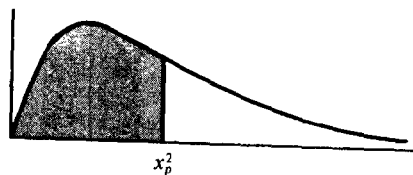
t - 分布的下侧分位数
(其中 ν 为自由度)



ν	$t_{.995}$	$t_{.99}$	$t_{.975}$	$t_{.95}$	$t_{.90}$	$t_{.80}$	$t_{.75}$	$t_{.70}$	$t_{.60}$	$t_{.55}$
1	63.66	31.82	12.71	6.31	3.08	1.376	1.000	.727	.325	.158
2	9.92	6.96	4.30	2.92	1.89	1.061	.816	.617	.289	.142
3	5.84	4.54	3.18	2.35	1.64	.978	.765	.584	.277	.137
4	4.60	3.75	2.78	2.13	1.53	.941	.741	.569	.271	.134
5	4.03	3.36	2.57	2.02	1.48	.920	.727	.559	.267	.132
6	3.71	3.14	2.45	1.94	1.44	.906	.718	.553	.265	.131
7	3.50	3.00	2.36	1.90	1.42	.896	.711	.549	.263	.130
8	3.36	2.90	2.31	1.86	1.40	.889	.706	.546	.262	.130
9	3.25	2.82	2.26	1.83	1.38	.883	.703	.543	.261	.129
10	3.17	2.76	2.23	1.81	1.37	.879	.700	.542	.260	.129
11	3.11	2.72	2.20	1.80	1.36	.876	.697	.540	.260	.129
12	3.06	2.68	2.18	1.78	1.36	.873	.695	.539	.259	.128
13	3.01	2.65	2.16	1.77	1.35	.870	.694	.538	.259	.128
14	2.98	2.62	2.14	1.76	1.34	.868	.692	.537	.258	.128
15	2.95	2.60	2.13	1.75	1.34	.866	.691	.536	.258	.128
16	2.92	2.58	2.12	1.75	1.34	.865	.690	.535	.258	.128
17	2.90	2.57	2.11	1.74	1.33	.863	.689	.534	.257	.128
18	2.88	2.55	2.10	1.73	1.33	.862	.688	.534	.257	.127
19	2.86	2.54	2.09	1.73	1.33	.861	.688	.533	.257	.127
20	2.84	2.53	2.09	1.72	1.32	.860	.687	.533	.257	.127
21	2.83	2.52	2.08	1.72	1.32	.859	.686	.532	.257	.127
22	2.82	2.51	2.07	1.72	1.32	.858	.686	.532	.256	.127
23	2.81	2.50	2.07	1.71	1.32	.858	.685	.532	.256	.127
24	2.80	2.49	2.06	1.71	1.32	.857	.685	.531	.256	.127
25	2.79	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
26	2.78	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
27	2.77	2.47	2.05	1.70	1.31	.855	.684	.531	.256	.127
28	2.76	2.47	2.05	1.70	1.31	.855	.683	.530	.256	.127
29	2.76	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
30	2.75	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
40	2.70	2.42	2.02	1.68	1.30	.851	.681	.529	.255	.126
60	2.66	2.39	2.00	1.67	1.30	.848	.679	.527	.254	.126
120	2.62	2.36	1.98	1.66	1.29	.845	.677	.526	.254	.126
∞	2.58	2.33	1.96	1.645	1.28	.842	.674	.524	.253	.126

附录IV

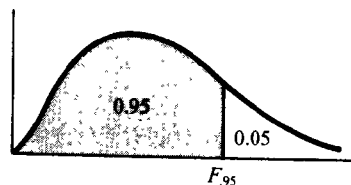
χ^2 分布的下侧分位数
(其中 ν 为自由度)



ν	$\chi^2_{.995}$	$\chi^2_{.99}$	$\chi^2_{.975}$	$\chi^2_{.95}$	$\chi^2_{.90}$	$\chi^2_{.75}$	$\chi^2_{.50}$	$\chi^2_{.25}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.01}$	$\chi^2_{.005}$
1	7.88	6.63	5.02	3.84	2.71	1.32	.455	.102	.0158	.0039	.0010	.0002	.0000
2	10.6	9.21	7.38	5.99	4.61	2.77	1.39	.575	.211	.103	.0506	.0201	.0100
3	12.8	11.3	9.35	7.81	6.25	4.11	2.37	1.21	.584	.352	.216	.115	.072
4	14.9	13.3	11.1	9.49	7.78	5.39	3.36	1.92	1.06	.711	.484	.297	.207
5	16.7	15.1	12.8	11.1	9.24	6.63	4.35	2.67	1.61	1.15	.831	.554	.412
6	18.5	16.8	14.4	12.6	10.6	7.84	5.35	3.45	2.20	1.64	1.24	.872	.676
7	20.3	18.5	16.0	14.1	12.0	9.04	6.35	4.25	2.83	2.17	1.69	1.24	.989
8	22.0	20.1	17.5	15.5	13.4	10.2	7.34	5.07	3.49	2.73	2.18	1.65	1.34
9	23.6	21.7	19.0	16.9	14.7	11.4	8.34	5.90	4.17	3.33	2.70	2.09	1.73
10	25.2	23.2	20.5	18.3	16.0	12.5	9.34	6.74	4.87	3.94	3.25	2.56	2.16
11	26.8	24.7	21.9	19.7	17.3	13.7	10.3	7.58	5.58	4.57	3.82	3.05	2.60
12	28.3	26.2	23.3	21.0	18.5	14.8	11.3	8.44	6.30	5.23	4.40	3.57	3.07
13	29.8	27.7	24.7	22.4	19.8	16.0	12.3	9.30	7.04	5.89	5.01	4.11	3.57
14	31.3	29.1	26.1	23.7	21.1	17.1	13.3	10.2	7.79	6.57	5.63	4.66	4.07
15	32.8	30.6	27.5	25.0	22.3	18.2	14.3	11.0	8.55	7.26	6.26	5.23	4.60
16	34.3	32.0	28.8	26.3	23.5	19.4	15.3	11.9	9.31	7.96	6.91	5.81	5.14
17	35.7	33.4	30.2	27.6	24.8	20.5	16.3	12.8	10.1	8.67	7.56	6.41	5.70
18	37.2	34.8	31.5	28.9	26.0	21.6	17.3	13.7	10.9	9.39	8.23	7.01	6.26
19	38.6	36.2	32.9	30.1	27.2	22.7	18.3	14.6	11.7	10.1	8.91	7.63	6.84
20	40.0	37.6	34.2	31.4	28.4	23.8	19.3	15.5	12.4	10.9	9.59	8.26	7.43
21	41.4	38.9	35.5	32.7	29.6	24.9	20.3	16.3	13.2	11.6	10.3	8.90	8.03
22	42.8	40.3	36.8	33.9	30.8	26.0	21.3	17.2	14.0	12.3	11.0	9.54	8.64
23	44.2	41.6	38.1	35.2	32.0	27.1	22.3	18.1	14.8	13.1	11.7	10.2	9.26
24	45.6	43.0	39.4	36.4	33.2	28.2	23.3	19.0	15.7	13.8	12.4	10.9	9.89
25	46.9	44.3	40.6	37.7	34.4	29.3	24.3	19.9	16.5	14.6	13.1	11.5	10.5
26	48.3	45.6	41.9	38.9	35.6	30.4	25.3	20.8	17.3	15.4	13.8	12.2	11.2
27	49.6	47.0	43.2	40.1	36.7	31.5	26.3	21.7	18.1	16.2	14.6	12.9	11.8
28	51.0	48.3	44.5	41.3	37.9	32.6	27.3	22.7	18.9	16.9	15.3	13.6	12.5
29	52.3	49.6	45.7	42.6	39.1	33.7	28.3	23.6	19.8	17.7	16.0	14.3	13.1
30	53.7	50.9	47.0	43.8	40.3	34.8	29.3	24.5	20.6	18.5	16.8	15.0	13.8
40	66.8	63.7	59.3	55.8	51.8	46.6	39.3	33.7	29.1	26.5	24.4	22.2	20.7
50	79.5	76.2	71.4	67.5	63.2	56.3	49.3	42.9	37.7	34.8	32.4	29.7	28.0
60	92.0	88.4	83.3	79.1	74.4	67.0	59.3	52.3	46.5	43.2	40.5	37.5	35.5
70	104.2	100.4	95.0	90.5	85.5	77.6	69.3	61.7	55.3	51.7	48.8	45.4	43.3
80	116.3	112.3	106.6	101.9	96.6	88.1	79.3	71.1	64.3	60.4	57.2	53.5	51.2
90	128.3	124.1	118.1	113.1	107.6	98.6	89.3	80.6	73.3	69.1	65.6	61.8	59.2
100	140.2	135.8	129.6	124.3	118.5	109.1	99.3	90.1	82.4	77.9	74.2	70.1	67.3

附录 V

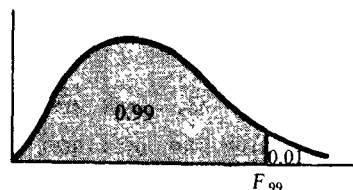
F - 分布的 95% 的下侧分位数
(其中 ν_1 为分子自由度, ν_2 为分母自由度)



$\nu_1 \backslash \nu_2$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

附录 VI

F - 分布的 99% 的下侧分位数
(其中 ν_1 为分子自由度, ν_2 为分母自由度)



$\nu_1 \backslash \nu_2$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	5000	5403	5625	5764	5859	5928	5981	6023	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	62.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.82	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

附录Ⅵ

常用对数表

N											比例部分								
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
N	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

续表

N											比例部分								
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	2	3	4	5	6	7	7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	4	5	5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	5
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
N	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

续表

N	0	1	2	3	4	5	6	7	8	9	比例部分								
											1	2	3	4	5	6	7	8	9
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0	1	1	2	2	3	3	4	4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	3	4
N	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

附录Ⅷ

$e^{-\lambda}$ 值 ($0 < \lambda < 1$)										
λ	0	1	2	3	4	5	6	7	8	9
0.0	1.0000	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
0.1	.9048	.8958	.8869	.8781	.8694	.8607	.8521	.8437	.8353	.8270
0.2	.8187	.8106	.8025	.7945	.7866	.7788	.7711	.7634	.7558	.7483
0.3	.7408	.7334	.7261	.7189	.7118	.7047	.6977	.6907	.6839	.6771
0.4	.6703	.6636	.6570	.6505	.6440	.6376	.6313	.6250	.6188	.6126
0.5	.6065	.6005	.5945	.5886	.5827	.5770	.5712	.5655	.5599	.5543
0.6	.5488	.5434	.5379	.5326	.5273	.5220	.5169	.5117	.5066	.5016
0.7	.4966	.4916	.4868	.4819	.4771	.4724	.4677	.4630	.4584	.4538
0.8	.4493	.4449	.4404	.4360	.4317	.4274	.4232	.4190	.4148	.4107
0.9	.4066	.4025	.3985	.3946	.3906	.3867	.3829	.3791	.3753	.3716
$(\lambda = 1, 2, 3, \dots, 10)$										
λ	1	2	3	4	5	6	7	8	9	10
$e^{-\lambda}$.36788	.13534	.04979	.01832	.006738	.002479	.000912	.000335	.000123	.000045

附录Ⅹ

随机数表

51772	74640	42331	29044	46621	62898	93582	04186	19640	87056
24033	23491	83587	06568	21960	21387	76105	10863	97453	90581
45939	60173	52078	25424	11645	55870	56974	37428	93507	94271
30586	02133	75797	45406	31041	86707	12973	17169	88116	42187
03585	79353	81938	82322	96799	85659	36081	50884	14070	74950
64937	03355	95863	20790	65304	55189	00745	65253	11822	15804
15630	64759	51135	98527	62586	41889	25439	88036	24034	67283
09448	56301	57683	30277	94623	85418	68829	06652	41982	49159
21631	91157	77331	60710	52290	16835	48653	71590	16159	14676
91097	17480	29414	06829	87843	28195	27279	47152	35683	47280
50532	25496	95652	42457	73547	76552	50020	24819	52984	76168
07136	40876	79971	54195	25708	51817	36732	72484	94923	75936
27989	64728	10744	08396	56242	90985	28868	99431	50995	20507
85184	73949	36601	46253	00477	25234	09908	36574	72139	70185
54398	21154	97810	36764	32869	11785	55261	59009	38714	38723
65544	34371	09591	07839	58892	92843	72828	91341	84821	63886
08263	65952	85762	64236	39238	18776	84303	99247	46149	03229
39817	67906	48236	16057	81812	15815	63700	85915	19219	45943
62257	04077	79443	95203	02479	30763	92486	54083	23631	05825
53298	90276	62545	21944	16530	03878	07516	95715	02526	33537